Multiple Imputation Methods for Nonignorable Nonresponse, Adaptive Survey Design, and Dissemination of Synthetic Geographies

by

Thais Viana Paiva

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Jerome Reiter, Supervisor

_____
Alan Gelfand

_____
Mine Çetinkaya-Rundel

_____
D. Sunshine Hillygus

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2014

Abstract

Multiple Imputation Methods for Nonignorable Nonresponse, Adaptive Survey Design, and Dissemination of Synthetic Geographies

by

Thais Viana Paiva

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Jerome Reiter, Supervisor

_____
Alan Gelfand

_____
Mine Çetinkaya-Rundel

_____
D. Sunshine Hillygus

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2014

Abstract

This thesis presents methods for multiple imputation that can be applied to missing data and data with confidential variables. Imputation is useful for missing data because it results in a data set that can be analyzed with complete data statistical methods. The missing data are filled in by values generated from a model fit to the observed data. The model specification will depend on the observed data pattern and the missing data mechanism. For example, when the reason why the data is missing is related to the outcome of interest, that is nonignorable missingness, we need to alter the model fit to the observed data to generate the imputed values from a different distribution. Imputation is also used for generating synthetic values for data sets with disclosure restrictions. Since the synthetic values are not actual observations, they can be released for statistical analysis. The interest is in fitting a model that approximates well the relationships in the original data, keeping the utility of the synthetic data, while preserving the confidentiality of the original data. We consider applications of these methods to data from social sciences and epidemiology.

The first method is for imputation of multivariate continuous data with nonignorable missingness. Regular imputation methods have been used to deal with nonresponse in several types of survey data. However, in some of these studies, the assumption of missing at random is not valid since the probability of missing depends on the response variable. We propose an imputation method for multivariate data sets when there is nonignorable missingness. We fit a truncated Dirichlet process

mixture of multivariate normals to the observed data under a Bayesian framework to provide flexibility. With the posterior samples from the mixture model, an analyst can alter the estimated distribution to obtain imputed data under different scenarios. To facilitate that, I developed an R application that allows the user to alter the values of the mixture parameters and visualize the imputation results automatically. I demonstrate this process of sensitivity analysis with an application to the Colombian Annual Manufacturing Survey. I also include a simulation study to show that the correct complete data distribution can be recovered if the true missing data mechanism is known, thus validating that the method can be meaningfully interpreted to do sensitivity analysis.

The second method uses the imputation techniques for nonignorable missingness to implement a procedure for adaptive design in surveys. Specifically, I develop a procedure that agencies can use to evaluate whether or not it is effective to stop data collection. This decision is based on utility measures to compare the data collected so far with potential follow-up samples. The options are assessed by imputation of the nonrespondents under different missingness scenarios considered by the analyst. The variation in the utility measures is compared to the cost induced by the follow-up sample sizes. We apply the proposed method to the 2007 U.S. Census of Manufactures.

The third method is for imputation of confidential data sets with spatial locations using disease mapping models. We consider data that include fine geographic information, such as census tract or street block identifiers. This type of data can be difficult to release as public use files, since fine geography provides information that ill-intentioned data users can use to identify individuals. We propose to release data with simulated geographies, so as to enable spatial analyses while reducing disclosure risks. We fit disease mapping models that predict areal-level counts from attributes in the file, and sample new locations based on the estimated models. I illustrate

this approach using data on causes of death in North Carolina, including evaluations of the disclosure risks and analytic validity that can result from releasing synthetic geographies.

To my family and Marcos, my greatest supporters

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

$N(\boldsymbol{\mu}, \Sigma)$	Multivariate normal distribution, with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.
$diag(\cdot)$	A diagonal matrix.

Abbreviations

MCAR	Missing Completely at Random.
MAR	Missing at Random.
MNAR	Missing Not at Random.
MI	Multiple Imputation.
MICE	Multivariate Imputation by Chained Equations.
NIMC	Nonignorable missingness Imputation for Multivariate Continuous data application.
MCMC	Markov Chain Monte Carlo.
MAP	Maximum a Posteriori.
GAM	Generalized Additive Models.
CAR	Conditionally Autoregressive.

Acknowledgements

I would like to thank everyone in the Department of Statistical Science for making every part of this possible. First, I want to thank my advisor, Jerry Reiter, for all his guidance, support and dedication over the past four years. He is a great inspiration for me, as a researcher, mentor and teacher. I also would like to thank Alan Gelfand and Mine Çetinkaya-Rundel for their helpful comments and guidance in our collaborations. I thank the faculty for the knowledge shared, and the staff for the support provided. Thanks to Quanli Wang, Hang Kim and to my collaborators at the U.S. Census Bureau. This work was supported by a grant from the National Science Foundation NSF(SES-1131897). Support for this research at the Triangle Census RDC from NSF(ITR-0427889) is also gratefully acknowledged.

I am really grateful for the inestimable help of all my friends. Thanks to my classmates Mary Beth, Maria, Tsuyoshi, Nick, Tommy, Tim and Daniel, for sharing moments of hard work and great fun throughout our journey. Thanks to Nicole, Jacopo, David, Monika and Christoph, for being amazing friends at all times. Thanks to Andrew and Jared, for being selflessly and helpfully smart. Thanks to my Brazilian friends Danilo, Vinícius, Fernando and Bruno, for helping make my life away from home easier. I want to thank everyone that participated in this process in any way: listening patiently to my stories and complaints, helping me on complex and simple statistical problems, and giving me rides even when it involved carrying heavy things. I am thankful for all the pleasant lunches, deep conversations and funny talks.

I would like to thank my family and friends in Brazil for the continuous support. Thanks to Renato, Erica, Marcos, Aline, Márcia, and many others from the Federal University of Minas Gerais, UFMG, for all the contribution to my academic career. They have been encouraging friends and helpful colleagues since my undergraduate years, through my time away during my PhD, and surely upon my return. Thanks to all my dear friends, cousins, aunts and uncles for always being present somehow and making me feel loved. Thanks to my grandparents, for being my greatest examples of strength and faith.

I am immensely thankful for the support of my parents, Alcy and Beatriz, and my brother, Leandro. They are my greatest encouragers and spared no effort to help me chase my goals. Lastly and most importantly, I want to thank my best friend and future husband, Marcos, for always being there for me. I am really grateful for having such an understanding and supporting partner, who has had an essential role in this achievement.

1

Introduction

Missing data are a common problem in many data sets, especially sample survey data. It can happen for different reasons and in diverse levels. For example, missing data can happen when an individual refuses to answer some questions in a survey interview or drops out of a longitudinal study. Missing data can be classified as two general types: item nonresponse and unit nonresponse. We say there is item nonresponse when a subject refuses to answer or skips one or more survey questions. In this case, some variables are recorded for that individual, but others are missing. We say there is unit nonresponse when the subject does not respond to the survey at all. In this case, none of the variables are available for that individual.

The simplest approach to deal with missing data is to use only the available observations and throw away the missing ones, called listwise deletion. This may be a reasonable solution when the proportion of missing data is small, for example, less than one percent. When this proportion is higher, however, deleting the incomplete cases not only increases the uncertainty around the estimates, but it can also lead to bias when the respondents are systematically different from the nonrespondents. For this reason, there is interest in developing methods for statistical analysis with

missing data.

There are many tools for dealing with missing data (Rubin, 1976; Brick and Kalton, 1996; Little and Rubin, 2002; Daniels and Hogan, 2008). Common ones include weighting methods, where the estimators are modified with individual weights to adjust for nonresponse; model-based methods, in which the analyst defines a model for the observed data and makes inference from the resulting likelihood; and imputation procedures, where the missing values are filled in to result in a data set that can be analyzed with complete data methods. Because of this potential of using complete data methods, imputation procedures are an attractive approach for statistical analysis with missing data. The missing data can be imputed by selecting observed values from similar observations such as hot deck imputation (Kalton and Kasprzyk, 1986), by using a deterministic function of the observed variables such as a conditional mean, or by sampling from a posterior distribution under a Bayesian approach. Rubin (1987) argues that the last approach is best among these, and proposes using multiple imputation, i.e., generating more than one value to replace the missing data. The advantage of this method is that it can preserve the joint probability distribution of the variables and appropriately propagate the uncertainty due to the missing data in inferences.

Multiple imputation approaches can also be used to solve a different problem: generate synthetic data for confidential data sets (Rubin, 1993; Little, 1993b; Raghunathan et al., 2003; Reiter, 2003, 2004; Reiter and Raghunathan, 2007; Wang and Reiter, 2012). In this case, the values that need to be imputed are not missing. Instead, they are sensitive information that cannot be released due to ethical or legal disclosure limitations. The agency can replace the sensitive values with draws from statistical models, creating multiply-imputed synthetic data sets. The challenge to the agency is to release data that respect the individuals' confidentiality while maintaining its utility.

This thesis presents some methods for multiple imputation that can be be applied to confidential and missing data problems. The thesis is organized in four chapters. In Chapter 2, I present an approach to impute multivariate continuous data when there is nonignorable unit nonresponse, i.e., the probability of missing depends on the variables of interest. In nonignorable missing data, the respondents and nonrespondents have different distributions. Since there is no data to estimate the latter, I propose to fit a model to the observed data and alter its parameters to create possible distributions for the missing data. I fit a truncated Dirichlet process prior to a mixture of multivariate normals to allow for flexibility when estimating the observed data distribution. To propose distributions for the nonrespondents, the mixture probabilities are altered, keeping fixed the location and scale parameters. An analyst evaluates different scenarios for the missingness pattern through sensitivity analysis. To facilitate this step, I developed an R application that enables the analyst to set new mixture probabilities and automatically generates data from the specified distribution. I demonstrate this method with an application to the Colombian Annual Manufacturing Survey. I also present the results of a simulation study to show that the correct complete data distribution can be recovered if the true missing data mechanism is known.

In Chapter 3, I use the imputation method for nonignorable missingness to implement a procedure for adaptive design in surveys. Specifically, I develop a procedure that agencies can use to evaluate whether or not it is effective to stop data collection and impute values for the nonrespondents, or to collect follow-up samples. This decision is based on the impact of possible missingness scenarios on the inference results, while considering the costs implied of collecting more data. I propose some utility measures to compare different imputation scenarios and guide the analyst's decision to stop data collection or not. I present the results of applying this method to the 2007 U.S. Census of Manufactures.

In Chapter 4, I present an approach to generate synthetic spatial locations using disease mapping models. These methods are useful when data include fine geographic information, such as census tract or street block identifiers, that cannot be released for public use. I propose to release data with simulated geographies, so as to enable spatial analyses while reducing disclosure risks. The basic idea is to fit disease mapping models to estimate the data occurrence intensity in the area over a specified grid. Then, the agency samples synthetic locations based on the estimated model. The size of the grid controls how well the original data distribution is approximated, and consequently the disclosure risk. I present some measures to assess the disclosure risk and the data utility of a data set with synthetic spatial locations. I apply the proposed approach to a data set with mortality records in North Carolina (NC), generating synthetic locations and evaluating their utility and risk.

In the remainder of this chapter, I provide some background information on some of the models and common definitions that are used throughout this thesis. In Section 1.1, I describe some different missing data mechanisms that are the motivation for the methods from Chapters 2 and 3. In Section 1.2, I introduce the concept of multiple imputation and some combining rules that are used for inference in the remaining chapters. Finally, in Section 1.3, I briefly review areal spatial models that are used for imputation of synthetic locations in Chapter 4.

1.1 Missing Data Mechanisms

When dealing with incomplete data, it is essential to consider the reasons that lead to missing observations. We refer to this as the missing data mechanism, following the concepts formalized by Rubin (1976). In this section, I briefly review such mechanisms, which are especially important for the imputation models in Chapters 2 and 3.

Let $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$ denote the $n \times p$ matrix of complete data, where $\boldsymbol{Y}_{obs}$ is

the observed data and $\boldsymbol{Y}_{mis}$ is the missing data. Let $\boldsymbol{\theta}$ denote the vector of the parameters for the model for $\boldsymbol{Y}$. Denote the missingness indicator by the matrix $\boldsymbol{R}$, also of size $n \times p$, where $r_{ij} = 1$ if, for individual i, the variable j is missing, and $r_{ij} = 0$ if it is observed. The definitions of missing data mechanisms depend on the relationship between the full data joint distribution, $f(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\theta})$, and the conditional probability distribution of the missingness indicator given the full data, $f(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{\theta})$.

When the reason for missing data does not depend on the response variables $\boldsymbol{Y}$, missing or observed, we say the data are *missing completely at random* (MCAR). This corresponds to

$$f(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{\theta}) \;\; = \;\; f(\boldsymbol{r}|\boldsymbol{\theta}), \tag{1.1}$$

that is, the missingness is unrelated to any variable. This implies that $\boldsymbol{Y}_{obs}$ and $\boldsymbol{Y}_{mis}$ have the same distribution, since the full data joint distribution can be factored as $f(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{\theta}) f(\boldsymbol{r}|\boldsymbol{\theta})$. Therefore, one can make valid inference about the distribution of $\boldsymbol{Y}$ based only on the respondents. For example, if an interviewer selects a random sample of the survey respondents to answer to an specific question, the ones that did not respond are MCAR.

A less restrictive case is when the reason for missing data does not depend on the missing responses $\boldsymbol{Y}_{mis}$, given the observed responses $\boldsymbol{Y}_{obs}$. In this case, we say the data are *missing at random* (MAR) and write

$$f(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{\theta}) \;\; = \;\; f(\boldsymbol{r}|\boldsymbol{y}_{obs}, \boldsymbol{\theta}). \tag{1.2}$$

Here, the probability of missing is explained by the part of the data that is observed. For example, we have MAR missingness if women are more likely than men to respond to a question about income and gender is an observed variable in the survey.

When the missing data mechanism is MAR and the parameters $\boldsymbol{\theta}$ can be partitioned into independent parts to index separately the full data model and the missing

data mechanism, we say the missing data mechanism is ignorable for posterior inference (Rubin, 1976; Little and Rubin, 1987). Under this condition, we can decompose the parameters as $\boldsymbol{\theta} = (\boldsymbol{\eta}, \boldsymbol{\phi})$, and the full data model and missingness model are written as $f(\boldsymbol{y}|\boldsymbol{\eta})$ and $f(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{\phi})$ respectively. Moreover, the parameters are *a priori* independent, that is, $p(\boldsymbol{\eta}, \boldsymbol{\phi}) = p(\boldsymbol{\eta})p(\boldsymbol{\phi})$. This facilitates inference about $\boldsymbol{\eta}$, since it can be done based on the posterior distribution $f(\boldsymbol{\eta}|\boldsymbol{Y}_{obs})$ with no need to specify the missing data mechanism.

When the reason for missing data depends on the missing responses, even conditioning on the observed data, we say the data are *missing not at random* (MNAR). That is,

$$f(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{\theta}) \quad = \quad f(\boldsymbol{r}|\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}, \boldsymbol{\theta}). \tag{1.3}$$

In this case, the value of the missing variable is related to the reason it is missing, and we say the missingness mechanism is nonignorable. For example, income is MNAR if higher income respondents are more likely to refuse to answer to a question about income.

It is impossible to verify if the data are MNAR with the data that was observed (Molenberghs and Kenward, 2007; Molenberghs et al., 2008). This is because, under MNAR, the distribution of $\boldsymbol{Y}$ cannot be identified for cases with missing values. Thus, to make some inference, it is necessary to specify a joint model of the responses and the missingness indicator (Little and Rubin, 2002; Daniels and Hogan, 2008). I describe some approaches for this purpose in Section 1.1.1.

Since assuming MNAR involves unverifiable assumptions, it is important to have a mechanism for sensitivity analysis. This involves the process of translating the prior beliefs about the missing data into the model, which is facilitated by prior distributions in a Bayesian framework, and evaluating the sensitivity of the results to different model specifications. The sensitivity analysis can be made in terms

of the sensitivity parameters that control the departure from MAR, or in terms of inferences with the full data (Rubin, 1977; Scharfstein et al., 2003; Molenberghs and Kenward, 2007; Daniels and Hogan, 2008; Fitzmaurice et al., 2008; Molenberghs, 2009). I present a tool for sensitivity analysis of multivariate continuous data in Chapter 2.

1.1.1 Models for nonignorable missingness

When dealing with missing data under MNAR, two common approaches are selection models and pattern-mixture models (Little and Rubin, 1987; Little, 1995, 2008). In both cases, we factor the full data joint distribution into marginal distributions of the response variables and the missing status. The difference is the conditional distribution of each term.

In selection models (Glynn et al., 1986; Rubin, 1987; Diggle and Kenward, 1994; Molenberghs et al., 1997), the full data joint distribution is factored into the marginal response model and the conditional distribution of the missingness given the response, such as

$$f(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\theta}) \quad = \quad f(\boldsymbol{y}|\boldsymbol{\theta}) f(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{\theta}). \tag{1.4}$$

The advantage of this approach is that the analyst specifies $f(\boldsymbol{y}|\boldsymbol{\theta})$, the full data response model, directly. The conditional distribution of the missingness indicator facilitates the generalization from ignorability to nonignorability and the association with the missing data mechanisms described previously. However, in these models, the missing data distribution is identified with parametric assumptions about the factorization terms in (1.4). This can be a problem, since the implied assumptions about the missing data mechanism can be hard to modify in a sensitivity analysis (Little, 1993a; Molenberghs et al., 1998). For example, consider a simple univariate case, where Y is assumed to follow a normal distribution, and the missing mechanism

is a logistic model linear on Y. In this case, the coefficient of Y in the logistic model for R serves as the sensitivity parameter. This coefficient has an identifiable estimate because of the normal distribution assumption for Y. However, it is not clear how one would meaningfully assess sensitivity to other assumptions about the missingness, as it is not obvious how to modify the model to encode other assumptions about the missing data (Daniels and Hogan, 2008, Chapter 8).

In pattern-mixture models (Glynn et al., 1993; Little, 1993a, 1994; Molenberghs et al., 1998; Thijs et al., 2002), the full data joint distribution is factored into the marginal distribution of the missing indicator and the conditional distribution of the response model given the missing indicator, such as

$$f(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\theta}) \quad = \quad f(\boldsymbol{y}|\boldsymbol{r}, \boldsymbol{\theta}) f(\boldsymbol{r}|\boldsymbol{\theta}). \tag{1.5}$$

The response model is a mixture denoted by

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{\boldsymbol{r}\in\{0,1\}} f(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\theta}) = \sum_{\boldsymbol{r}\in\{0,1\}} f(\boldsymbol{y}|\boldsymbol{r}, \boldsymbol{\theta})\, f(\boldsymbol{r}|\boldsymbol{\theta}). \tag{1.6}$$

The advantage of this approach is that it can facilitate the sensitivity analysis, since it is necessary to specify how distributions for the respondents and nonrespondents are different. This can be complicated in longitudinal studies with many stages being analyzed. However, it can also make it easier to identify the sensitivity parameters that are not identified by the observed data (Daniels and Hogan, 2008). We follow this approach and present a method to make the distinction between the distribution of the observed and missing data in Chapter 2.

1.2 Multiple Imputation

Imputation consists of filling in the missing observations to make a completed data set. The missing values can be replaced by deterministic values, such as marginal and

conditional means, or sampled from probability models fit to the data. Imputation is an attractive approach for dealing with missing data, since it allows analysts to run standard methods for statistical analysis of complete data. However, using ad hoc methods, such as imputing averages or regression predictions, can become problematic in a multivariate setup. For example, plugging in marginal means can alter the original correlations between the variables (Little and Rubin, 2002). Thus, simulating values from a imputation model is a better alternative to preserve the joint distribution of the variables.

Another issue with single imputation strategies, whether deterministic or model-based, is that analysis of the complete data effectively implies that the imputed value has no uncertainty. To account for this variability, we can use multiple imputation (MI) proposed by Rubin (1987). It consists of generating $m > 1$ versions of the complete data set, where the imputed values are sampled from their predictive distribution. Generally, it is enough to generate a small number of imputation, with m between 5-10, to yield efficient estimation (Rubin, 1987). With the m complete data sets, each version is analyzed separately with standard complete-data methods. The results are then combined with the appropriate rules that we describe in Section 1.2.1.

With the development of Markov Chain Monte Carlo (MCMC) methods, it became easier to implement multiple imputation (Schafer, 1997; Schafer and Olsen, 1998). This is done via data augmentation, where the imputed values are sampled within each iteration of the MCMC. Consider the full data matrix $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$. After specifying the full model, $f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}|\boldsymbol{\theta})$ and the prior distributions for $\boldsymbol{\theta}$, the imputed values are simulated from their predictive distribution via Gibbs sampling.

At iteration t, we sample

$$\boldsymbol{y}_{mis}^{(t)} \sim f(\boldsymbol{y}_{mis}|\boldsymbol{y}_{obs}, \boldsymbol{\theta}^{(t-1)}) \tag{1.7}$$

$$\boldsymbol{\theta}^{(t)} \sim f(\boldsymbol{\theta}|\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}^{(t)}). \tag{1.8}$$

After convergence, we select a subset of m samples sufficiently spread to provide approximate independence, creating proper multiple imputed data sets (Schafer, 1997).

Since the multiple imputed data sets can be analyzed with complete data methods, MI has been used to handle missing data and confidential data in many applications. A review of these applications and their extensions can be seen in Rubin (1996), Schafer (1997), Barnard and Meng (1999), Harel and Zhou (2007), and Reiter and Raghunathan (2007).

An alternative to joint modeling of $\boldsymbol{Y}$ is using fully conditional specification of each variable, also known as multivariate imputation by chained equations (MICE) (Van Buuren and Oudshoorn, 2000; Raghunathan et al., 2001). It is attractive when the multivariate joint distribution cannot be specified, for example, when dealing with mixed data. This advantage made MICE a popular method applied in many fields (van Buuren and Groothuis-Oudshoorn, 2011). However, one of MICE's drawbacks is that it can result in an improper joint model (Liu et al., 2014).

1.2.1 Inference for Multiple Imputation

Rubin (1987) proposed a set of combining rules for inference with data generated with multiple imputation. Let Q denote a quantity of interest of the population, such as mean or regression coefficient. For each completed data set generated with MI indexed by $i = 1, \ldots, m$, let q_i denote a point estimator of Q and v_i denote an

estimator of the variance of q_i. We need the following quantities for inference.

$$\bar{q}_m = \sum_{i=1}^{m} q_i/m \tag{1.9}$$

$$b_m = \sum_{i=1}^{m} (q_i - \bar{q}_m)^2/(m-1) \tag{1.10}$$

$$\bar{v}_m = \sum_{i=1}^{m} v_i/m \tag{1.11}$$

$$T_m = \bar{v}_m + \left(1 + \frac{1}{m}\right) b_m. \tag{1.12}$$

We use $\bar{q}_m$ to estimate Q, and T_m to estimate the variance of $\bar{q}_m$. With a modest value of m and large n, Q follows a Student's t distribution with degrees of freedom $\nu_m = (m-1)\left(1 + \bar{v}_m/[(1+1/m)b_m]\right)^2$. Barnard and Rubin (1999) present a modified formula for the degrees of freedom for small samples. Extensions of these rules for methods for multivariate analysis are described in Li et al. (1991), Meng and Rubin (1992), Raghunathan et al. (2001) and Reiter (2007).

Some variations of these rules have been proposed in the literature (Raghunathan et al., 2003; Reiter and Raghunathan, 2007), including the rules proposed by Reiter (2003) for partially synthetic data. These apply to data sets that include a mix of original observations and some imputed values that can replace, for example, some sensitive variables. Again, the interest is on estimating a quantity Q with estimator q_i and variance estimate v_i for each complete synthetic data set i, with $i = 1, \ldots, m$. In this case, we use the same quantities $\bar{q}_m$, b_m and $\bar{v}_m$ from equations (1.9)–(1.11). However, the variance estimator of $\bar{q}_m$ is given by

$$T_p = \bar{v}_m + \left(\frac{1}{m}\right) b_m. \tag{1.13}$$

Similarly, inference can be based on a t distribution with $\nu_m = (m-1)\left(1 + m\bar{v}_m/b_m\right)^2$

degrees of freedom. The difference between the two approaches is discussed in Reiter and Raghunathan (2007).

Multiple imputed data sets can also be used for Bayesian analysis. When the posterior distributions of the parameters of interest are approximately normal, the regular approach can be used: obtain the estimates from each synthetic data set, and then use the combining rules to obtain the results. However, the rules proposed by Rubin (1987) are not adequate when the posterior distributions depart from normality. Since Bayesian inference is an attractive approach for these cases, Zhou et al. (2010) demonstrate that a better alternative is to simulate posterior draws for each imputed data set and then combine all the draws. For each $i = 1, \ldots, m$, we simulate a large number S of values of Q, the parameter of interest, from its conditional distribution given the i-th imputed data set. The draws are combined to form a list of size mS. This posterior sample is then used for inference about Q.

1.3 Synthetic Spatial Locations

With the increasing availability of technologies to record geographical locations, geospatial data have become a frequent part of data collection and a rich source of information in various fields. However, when agencies collect microdata about individuals, they have to observe promises to protect confidentiality before releasing the data, especially when they include geographical locations. The agencies must protect the respondents' identities and sensitive attributes due to ethical and legal reasons. Thus, for some data to be released and used by researchers, the agencies need to reduce risks of disclosure of confidential information. Some common ways to reduce disclosure risks of data sets with linked spatial coordinates are described in the report "Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data" from the National Research Council (2007). Most agencies aggregate data by geographic units, such as counties or cities. This strategy preserves

the confidentiality of the individuals at the level of aggregation, but it does not allow for analyses on smaller areas. Other strategies for protecting geography include adding random noise to locations, e.g., Armstrong et al. (1999) and VanWey et al. (2005); or swapping individuals across locations, e.g., Zayatz (2007) and Young et al. (2009). The problem with these strategies is that they can affect the inference results of the released data.

To help preserve the data utility, Wang and Reiter (2012) propose releasing synthetic locations simulated from statistical models. In these models, the spatial locations are treated as outcome variables to be predicted from the other attributes on the file. This strategy can protect confidentiality, since the data released does not include actual observed locations. It also can preserve the associations in the original data, given that they are captured by the model, and allows statistical analysis at finer geographic level. We propose a similar approach in Chapter 4, where the spatial locations are generated from areal level spatial models that we describe next.

1.3.1 Areal level spatial models

With areal data, the study region is partitioned into a finite number, say G, of nonoverlapping areas. These geographical units can have a regular shape, for example, created by an artificial grid, or irregular shapes, such as census tracts or counties. The data available often consist of sums or averages of variables over these areas.

Areal data models are frequently used in epidemiological studies as a tool for disease mapping. In these models, the variables normally include the number of disease cases and population at risk at each area. Let C_i denote the counts of disease cases in area i, where i indexes the areas with $i = 1, \ldots, G$. Similarly, let E_i denote the expected number of cases in area i, where this number is calculated by applying an overall disease rate to the number of people at risk in each area denoted by P_i. This disease rate can be calculated via internal standardization as the overall

observed rate, $\bar{r} = \sum_i c_i / \sum_i P_i$, with $E_i = P_i \bar{r}$, or via external standardization when there is access to a table of disease rates.

After observing the counts c_i, we seek to calculate the true relative risk of disease in each area. This can be done by assuming that $C_i|\lambda_i \sim \text{Poisson}(E_i\lambda_i)$ and estimating the risk λ_i. This model can be expanded to include random effects through a hierarchical Bayesian model (Clayton and Kaldor, 1987; Besag et al., 1991; Clayton and Bernardinelli, 1992; Wakefield, 2007), allowing for spatial correlation between the areas. Thus, we have

$$C_i|\lambda_i \sim \text{Poisson}(E_i\lambda_i) \tag{1.14}$$

$$\log \lambda_i = X_i\boldsymbol{\beta} + \theta_i + \epsilon_i, \tag{1.15}$$

where X_i is a p-dimensional vector with covariate values for area i, and β contains the p coefficients for each explanatory covariate. The term θ_i is an area-specific spatial effect, and ϵ_i is an error term to capture extra variability in the Poisson model. Analysts typically use normal prior distributions for coefficients $\boldsymbol{\beta}$ and the error terms $\boldsymbol{\epsilon}$, such as:

$$\beta_k \sim N(0, \sigma_\beta^2) \quad \text{for } k = 1, \ldots, p \tag{1.16}$$

$$\epsilon_i \sim N(0, \sigma_\epsilon^2) \quad \text{for } i = 1, \ldots, G. \tag{1.17}$$

To allow for spatial association between neighboring areas, analysts can use an intrinsic CAR (conditionally autoregressive) prior distribution (Banerjee et al., 2004) for the spatial effects, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_G)$. This corresponds to the conditional distribution,

$$\theta_i|\boldsymbol{\theta}_{-i} \sim N\left(\bar{\theta}_i, \sigma_\theta^2/n_i\right), \tag{1.18}$$

where $\boldsymbol{\theta}_{-i}$ is the vector with θ_j for all $j \neq i$; $\bar{\theta}_i$ is the average of θ_j for all $j \sim i$, i.e., area j is neighbor of area i; and n_i is the number of neighbors of area i. The

neighborhood structure can be defined based on boundaries or distances between area centroids (Banerjee et al., 2004). The variance hyperparameters are specified with prior gamma distributions. We use an extended version of this model in Chapter 4, where we discuss more implementation details.

2

Imputation of multivariate continuous data with nonignorable missingness

2.1 Introduction

Missing data are common in nearly every field, arising for example from dropout in longitudinal studies, from study subjects' refusal to answer questions, and from failure to record values. As is well known, using only the cases with complete data for analysis can be inadequate. At best, it results in inefficiencies by sacrificing partially observed information, and at worst it results in bias when the reason for nonresponse is related to the outcome of interest. Thus, many researchers use alternatives to complete case analysis, including likelihood based and Bayesian methods (Little and Rubin, 1987; Rubin, 1976) and multiple imputation (Rubin, 1987). For overviews of methods for analyzing incomplete data, see for example Schafer (1997), Schafer and Graham (2002), Reiter and Raghunathan (2007), Daniels and Hogan (2008), and Graham (2012). In this chapter, we focus on multiple imputation for multivariate continuous data when there is nonignorable missingness. We assume familiarity with multiple imputation at the level described in Chapter 1.

Most of the methods for multiple imputation are developed under the assumption of missing at random (MAR), which means that the distribution of missingness does not depend on the missing data (Rubin, 1976). In many sample surveys and observational studies, the ignorability assumption is invalid, so that standard methods for multiple imputation can produce unreliable estimates. For that reason, it is useful and important to develop methods for missing not at random (MNAR) data. As examples, Greenlees et al. (1982) proposes an imputation method for nonignorable response mechanism by considering a regression model with censoring, and Diggle and Kenward (1994) present a modeling approach for longitudinal data sets with a nonignorable dropout process.

Two common approaches for dealing with nonignorable missingness are selection models and pattern mixture models (Little and Rubin, 1987; Little, 1995). In selection models, the joint distribution of the response variable and the missingness indicator is factored into the marginal response model and the conditional distribution of the missingness given the response. In pattern mixture models, it is the reverse: the joint distribution is factored into the marginal missing data mechanism and the conditional distribution of the response model given the missing status. As discussed in Chapter 1, parametric selection models have some limitations for analysis of sensitivity to different assumptions about the nonignorable missingness. The factorization of selection models does not separate the identified and unidentified parameters, making it difficult to check the assumptions with sensitivity analysis (Daniels and Hogan, 2008). The assumptions about the missingness mechanism are more easily formulated and interpreted in the factorization of the pattern mixture models. Therefore, we use a pattern mixture model approach, assuming that the respondents and nonrespondents have different distributions.

Fitting a pattern mixture model requires specifying the conditional distribution of the response variables given the missing data indicator. Since this distribution in-

cludes sensitivity parameters that depend on missing observations, the analyst needs to make identifying restrictions and assumptions about the missing data distribution (Little, 1993a). Under a Bayesian framework, these assumptions can be instantiated by specifying the prior distribution of the model parameters. As Greenlees et al. (1982, Section 2.2) point out, these assumptions can have a great impact on the inferences. Thus, it is important to compare inferences based on different prior specifications via sensitivity analysis (Daniels and Hogan, 2008, Chapter 9). An example of how to incorporate prior information and perform sensitivity analysis using pattern mixture models is proposed by Daniels and Hogan (2000) for longitudinal studies. They model the responses as a normal distribution for each dropout stage, yielding some restrictions on the location-scale parametrization for a monotone dropout.

We propose a method for handling nonignorable missingness when making multiple imputation of multivariate continuous data. To capture distributional features that this type of data may have, we use a mixture of multivariate normal distributions and a truncated Dirichlet process prior distribution. We fit the model to the set of observed data, resulting in an estimate of the respondents' distribution. When the data are missing not at random, we need a different distribution for the nonrespondents, following a pattern mixture model approach. To propose distributions for the nonrespondents, we alter the probabilities of the mixture components, keeping fixed the location and scale parameters. We inflate cluster probabilities to generate more points from that region, and deflate them to do the opposite, yielding a new distribution for the imputation. We then perform sensitivity analysis by examining different sets of altered probabilities. Since that is the main step of the imputation method, we developed an R application to implement the process of selecting new probabilities. The user selects new probabilities by setting the values on "sliders" for each components. The application automatically generates data from the specified distribution, and shows plots and summary statistics to be used for the sensitivity

analysis.

The remainder of the chapter is organized as follows. In Section 2.2, we describe the methodology used for fitting the model to the observed data (Section 2.2.1) and for generating imputed data sets under MNAR (Section 2.2.2). In Section 2.3, we demonstrate the method with an application to the Colombian Annual Manufacturing Survey. In Section 2.4, we present some simulation results to show that the correct complete data distribution can be recovered if the true missing data mechanism is known, thus validating that the method can be meaningfully interpreted to do sensitivity analysis. Finally, in Section 2.5, we discuss the advantages and limitations of the method.

2.2 Methodology

The full data consists of the response matrix $\boldsymbol{Y}$ and the missing data indicator $\boldsymbol{R}$. We assume that there are no covariates, that is, all variables are included in $\boldsymbol{Y}$. The response matrix $\boldsymbol{Y}$ has dimension $N \times p$, with rows $\boldsymbol{Y}_i = (y_{i1}, \ldots, y_{ip})$ for observation index $i = 1, \ldots, N$ and p outcome variables. The vector $\boldsymbol{R} = (r_1, \ldots, r_N)$ contains the missing data indicators for each observation in the sample. Let $r_i = 1$ if all the variables are missing for the i-th observation, i.e. the unit does not respond, and $r_i = 0$ otherwise. In this chapter, we use as respondents only the complete observations. For now, we consider unit nonresponse to be the source of nonignorable missingness. An extension of the model to deal with item nonresponse is discussed in Section 2.5.

In pattern mixture models, the full data distribution is factorized into a marginal distribution for the missing indicator and a conditional distribution of the response variables given the missing status. The factorization of the full data joint distribution can be written as

$$f(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{r}, \boldsymbol{\theta})\, f(\boldsymbol{r}|\boldsymbol{\theta}), \tag{2.1}$$

where $\boldsymbol{\theta}$ is the parameter space that can be partitioned into the corresponding parameters for each of the terms on the right side of (2.1).

Our interest is on the difference between the distribution of the respondents, $f(\boldsymbol{y}_i|r_i = 0, \boldsymbol{\theta})$, and the distribution of the nonrespondents, $f(\boldsymbol{y}_i|r_i = 1, \boldsymbol{\theta})$. Since there is no data to estimate the latter, we construct it based on the model fitted to the observed distribution. In Section 2.2.1, we describe the modeling approach used for estimating the distribution of the respondents. Then, in Section 2.2.2, we outline the steps for the imputation under MNAR and discuss the issues that can arise.

2.2.1 Mixture of multivariate normal distributions

With continuous survey data, we desire a flexible model able to capture arbitrary distributional features in the data. We use a mixture of normal distributions, since it can approximate any distribution if provided enough components. We use a Bayesian nonparametric prior (Ferguson, 1973, 1983; Escobar and West, 1995; West et al., 1994), since this class of models allows for more flexibility and better density estimation as reviewed recently by Müller and Mitra (2013). We use a truncated Dirichlet process prior with a stick-breaking representation and multivariate normal kernels (Sethuraman, 1994; Ishwaran and James, 2001).

Model specification with default prior

Let $\boldsymbol{Y}_{\text{com}} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)'$ denote n complete observations, each a p-dimensional variable, from individuals with $r_i = 0$. Each dimension of $\boldsymbol{Y}_{\text{com}}$ is standardized to facilitate modeling. Suppose each observation belongs to one of $K < \infty$ latent mixture components. Let $z_i \in 1, \ldots, K$ be the indicator of which component the i-th observation belongs to, with $i = 1, \ldots, n$. Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ denote the mixture probabilities of each component, with $\pi_k = P(z_i = k)$ for all i. Each component follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}_k$ and covariance matrix

Σ_k. The mixture model can be expressed as

$$\begin{aligned} \boldsymbol{y}_i|z_i, \boldsymbol{\mu}, \boldsymbol{\Sigma} &\sim N(\boldsymbol{\mu}_{z_i}, \Sigma_{z_i}) && (2.2) \\ z_i|\boldsymbol{\pi} &\sim \text{Multinomial}(\pi_1, \ldots, \pi_K). && (2.3) \end{aligned}$$

Integrating out the latent mixture components, the marginal mixture model is

$$p(\boldsymbol{y}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \, N(\boldsymbol{\mu}_k, \Sigma_k). \quad (2.4)$$

We follow the conjugate prior specification from Kim et al. (2014). For the normal parameters, we have

$$\begin{aligned} \boldsymbol{\mu}_k|\Sigma_k &\sim N(\boldsymbol{\mu}_0, h^{-1}\Sigma_k) && (2.5) \\ \Sigma_k &\sim \text{InverseWishart}(f, \Phi), && (2.6) \end{aligned}$$

where f is the degrees of freedom and $\Phi = diag(\phi_1, \ldots, \phi_p)$ with $\phi_j \sim \text{Gamma}(a_\phi, b_\phi)$ for $j = 1, \ldots, p$. Following the stick-breaking representation of a truncated Dirichlet process (Sethuraman, 1994; Ishwaran and James, 2001), the mixture probabilities are defined as

$$\begin{aligned} \pi_k &= v_k \textstyle\prod_{g<k}(1 - v_g) \quad \text{for } k = 1, \ldots, K && (2.7) \\ v_k &\sim \text{Beta}(1, \alpha) \quad \text{for } k = 1, \ldots, K-1; \; v_K = 1 && (2.8) \\ \alpha &\sim \text{Gamma}(a_\alpha, b_\alpha). && (2.9) \end{aligned}$$

For the hyperparameters, we also follow the discussion in Kim et al. (2014). We specify vague priors for the Gamma distributions with $a_\phi = b_\phi = 0.25$ to allow substantial prior mass at modest sized variances. We set $\boldsymbol{\mu}_0 = 0$, since the variables are standardized, $f = p + 1$ to ensure a proper posterior distribution, and $h = 1$ for convenience. We also specify vague priors for α by setting $a_\alpha = b_\alpha = 0.25$. The

stick-breaking representation in (2.7), with small values of α, encourages allocating the probabilities to the first few components, avoids overfitting and increases computational efficiency.

The choice of K depends on the dimensions of the data. We recommend starting with a large value, for example $K = 30$, and readjusting depending on the posterior results. If the model allocates a reasonable number of observations to all the clusters, it is prudent to increase K and fit the model again with more components. If the observations are allocated to less than K clusters, i.e., some are empty, then the choice of K is reasonable.

With the conjugate prior specification, we can obtain posterior samples using a Gibbs sampler algorithm (Ishwaran and James, 2001). For each component $k = 1, \ldots, K$, let $N_k = \sum_{i=1}^{n} \mathbb{1}_{(z_i=k)}$ be the number of observations at component k, and define $\bar{\boldsymbol{y}}_k = \sum_{\{i:z_i=k\}} \boldsymbol{y}_i/N_k$ and $S_k = \sum_{\{i:z_i=k\}} (\boldsymbol{y}_i - \bar{\boldsymbol{y}}_k)(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_k)'$. After initialization, the sampler iterates through the following steps:

1. For $k = 1, \ldots, K$, update $\boldsymbol{\mu}_k$ and Σ_k from the full conditionals,

$$\Sigma_k | \boldsymbol{y}, \boldsymbol{z} \sim \text{InverseWishart}(f_k, \Phi_k) \tag{2.10}$$

$$\boldsymbol{\mu}_k | \Sigma_k, \boldsymbol{y}, \boldsymbol{z} \sim N\left(\bar{\boldsymbol{\mu}}_k, \bar{\Sigma}_k\right), \tag{2.11}$$

   where $f_k = f + N_k$ and $\Phi_k = \Phi + S_k + (\bar{\boldsymbol{y}}_k - \boldsymbol{\mu}_0)(\bar{\boldsymbol{y}}_k - \boldsymbol{\mu}_0)'/(1/h + 1/N_k)$ are the parameters for sampling the covariance matrix; $\bar{\boldsymbol{\mu}}_k = (h\boldsymbol{\mu}_0 + N_k\bar{\boldsymbol{y}}_k)/(h + N_k)$ and $\bar{\Sigma}_k = \frac{1}{h+N_k}\Sigma_k$ are the parameters for sampling the mean.

2. For $k = 1, \ldots, K-1$, update v_k from the full conditional,

$$v_k | \boldsymbol{z}, \alpha \sim \text{Beta}\left(1 + N_k, \alpha + \sum_{g>k} N_g\right), \tag{2.12}$$

   and set $v_K = 1$. For $k = 1, \ldots, K$, set $\pi_k = v_k \prod_{g<k}(1 - v_g)$, following the specification from (2.7).

3. For $j = 1, \ldots, p$, update ϕ_j from the full conditional,

$$\phi_j|\boldsymbol{\Sigma} \;\sim\; \text{Gamma}\left(a_\phi + \frac{K(p+1)}{2}, b_\phi + \frac{1}{2}\sum_{k=1}^{K}\left(\Sigma_k^{-1}\right)_{(j,j)}\right), \tag{2.13}$$

where $\left(\Sigma_k^{-1}\right)_{(j,j)}$ is the j-th diagonal element of Σ_k^{-1}.

4. Update α from the full conditional,

$$\alpha|\boldsymbol{\pi} \;\sim\; \text{Gamma}\left(a_\alpha + K - 1, b_\alpha - \log \pi_K\right). \tag{2.14}$$

5. For $i = 1, \ldots, n$, update z_i from the full conditional,

$$z_i|\boldsymbol{y}_i, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \;\sim\; \text{Multinomial}(\pi_{i1}^*, \ldots, \pi_{iK}^*), \tag{2.15}$$

where $\pi_{ik}^* = \pi_k N(\boldsymbol{y}_i|\boldsymbol{\mu}_k, \Sigma_k) / \left\{\sum_{g=1}^{K} \pi_g N(\boldsymbol{y}_i|\boldsymbol{\mu}_g, \Sigma_g)\right\}$.

These steps are repeated for T iterations, after discarding the first iterations for burn-in. We denote the parameter samples at each iteration with the superscript (t), where $t = 1, \ldots, T$. The MCMC convergence is assessed by the traceplots of the parameters being sampled.

Model specification with fixed covariance matrices

For some $\boldsymbol{Y}_{\text{com}}$, the model may result in only a small number of occupied clusters; e.g., when the data are distributed homogeneously in compact regions. With a small number of occupied clusters, the options for altering the estimated distribution might be too limited for our imputation purposes. This limitation can also happen if the fitted clusters are close to each other in overlapping regions. In these situations, it might be infeasible to obtain different distributions for nonrespondents and respondents by adjusting the cluster probabilities. Our goal is to enable the user to have control over the imputation model and to sample from various missing data patterns

by changing the mixture probabilities. It might be advantageous for MNAR imputation purposes to change the prior specification to encourage more occupied and separated clusters, as this gives more options for customizing the nonrespondents' distribution.

One alternative is to force the model to fit more and tighter clusters. We suggest fixing the covariance matrices of all components to $\Sigma_k^{(t)} = \sigma I_p, \forall k$ and $\forall t$, with σ controlling the tightness of the clusters. Since Σ_k is fixed at all iterations, this model restriction is simple to implement by following the same MCMC steps as before and skipping the sampling of (2.10).

The choice of σ depends mostly on the range of the data. After standardizing the observations, we found that setting $\sigma < 1$ can improve the results by creating a reasonable number of occupied clusters that cover smaller regions. This results in a model that allows a more precise tuning of the mixture parameters and more flexibility for creating different imputation scenarios. We demonstrate how this can be an advantage in Section 2.3.

2.2.2 Imputation under MNAR

The posterior samples of the parameters $(\boldsymbol{\pi}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ from the model in (2.4) can provide a good estimate of the distribution of the observed part of the data, that is $f(\boldsymbol{y}_i | r_i = 0, \boldsymbol{\theta})$. Under a missing at random assumption, imputed data could be easily generated with the posterior predictive distribution. However, since we are dealing with observations that are missing not at random, we need to alter the estimated distribution to reflect the differences that are believed to affect the distribution of the nonrespondents, $f(\boldsymbol{y}_i | r_i = 1, \boldsymbol{\theta})$.

The distribution of the respondents can be altered by changing the mixture probabilities, locations or covariance structure from the estimated model. Changing the location and scale of the clusters require more prior information from the specialist

about where and how the nonrespondents should be located. The number of parameters to specify also increases quickly depending on the dimension of the data and the number of occupied components. On the other hand, specifying new probabilities for the components is a lower dimensional problem, especially with few occupied clusters, and perhaps more intuitive for the user, since it is based on what has been observed. For these reasons, we will focus on altering the distribution by changing the probabilities.

With the prior specification from (2.7)–(2.9), the probabilities tend to be allocated to the first components and decay fast. Thus, if the specified K is large enough, we expect to see some empty components. We denote by $\tilde{K}^{(t)} \leq K$ the number of occupied clusters at the t-th posterior sample, where $\tilde{K}^{(t)} = \sum_{k=1}^{K} \mathbb{1}_{\left(N_k^{(t)}>0\right)}$. We potentially change only the probabilities for the clusters with $N_k^{(t)} > 0$; that is, we leave the probabilities at zero for all empty clusters.

With new probabilities $\boldsymbol{\pi}^*$, we proceed to the imputation by generating data from the posterior predictive distributions using the samples of $\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\Sigma}^{(t)}$. Let n_{mis} denote the number of missing observations that will be imputed. For $i = 1, \ldots, n_{\text{mis}}$, select a cluster by sampling from

$$z_i|\boldsymbol{\pi}^* \quad \sim \quad \text{Multinomial}(\pi_1^*, \ldots, \pi_K^*). \tag{2.16}$$

Given z_i and posterior sample (t), generate the imputed response vector from

$$\tilde{\boldsymbol{y}}_i|z_i, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)} \quad \sim \quad N(\boldsymbol{\mu}_{z_i}^{(t)}, \Sigma_{z_i}^{(t)}), \tag{2.17}$$

where $\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\Sigma}^{(t)}$ are the posterior samples from the fitted model. An alternative to sampling from the predictive distribution is to use a modified hot-deck imputation: select a component with probabilities $\boldsymbol{\pi}^*$, and sample the imputed values randomly from the set of observations in that cluster. This simplified approach limits the imputed values to already observed responses, which respects the empirical association

among the variables. However, it can underestimate the variability in the imputed values and in the results of multiple imputation inference.

Choosing $\boldsymbol{\pi}^*$ determines which components the nonrespondents are more likely to belong to, and specifies the new pattern for the imputed data. The values in $\boldsymbol{\pi}^*$ are chosen by analyzing the posterior summary of the clusters, and adjusting the values to obtain the desired pattern for the imputed values. For example, analysts can set $\pi_k^* = 0$ for clusters in regions that should not have imputed data according to their beliefs about the the nonignorable missingness. On the other hand, analysts can increase the probability of clusters where they believe there should be more imputed data than what was observed. The closer the probabilities $\boldsymbol{\pi}^*$ remain to the posterior samples of $\boldsymbol{\pi}$, the closer the data are to MAR.

Specifying the entire vector of probabilities might be complicated in some situations, particularly with large p. Therefore, we recommend starting with and updating the estimated $\boldsymbol{\pi}^{(t)}$. We discuss how to choose t in Section 2.2.2. To facilitate the process of setting $\boldsymbol{\pi}^*$, we developed the NIMC (Nonignorable missingness Imputation for Multivariate Continuous data) tool. The NIMC is an R (R Core Team, 2013) application developed with the `shiny` package (RStudio and Inc., 2014), where the probabilities can be altered by setting values on sliders for each component. The sliders start with the estimated probabilities of $\boldsymbol{\pi}^{(t)}$ for a given iteration t as default values, and the user selects the factors to be applied to each probability. The slider values are renormed to sum to one so the resulting mixture model in (2.4) is a proper density. Using the renormed probabilities, the application automatically generates synthetic data following the model described in (2.16)–(2.17), so that the analyst can change $\boldsymbol{\pi}^*$ to get the desired distributions.

The application presents some tabs with the results of the imputation. The main tab includes pairwise scatterplots of the imputed and observed data, and the 95% quantile ellipses of the fitted clusters. We include the plot with the log transformed

and standardized data, on the scale in which the model was fit, and the plot with the data on its original scale. For the latter, the user has an option to choose between the raw data and the data with the log transformation to help visualize skewed distributions. To help with the visual aspect of the scenario assessment, a second tab includes summary statistics for the observed and imputed data separately, and the merged completed data set. The summary statistics are also calculated for the standardized and original data. The user can select the summary statistics, depending on the variables being analyzed. These measures, in combination with the scatterplots, provide different perspectives to help the analyst make decisions about the sensitivity analysis. As a baseline, we include correlations between the response variables and descriptive statistics for the marginal variables, with quantiles, means, minimum values and maximum values. Finally, when the imputation results are satisfactory, the completed data set on the original scale can be visualized and downloaded in the third tab.

In summary, these are the steps to follow to apply our method for imputation under MNAR:

1. Fit the mixture model to the complete observations as described in Section 2.2.1.

2. After running the MCMC long enough to achieve convergence, sort the samples based on a rank for the components as explained below.

3. Select the subset of posterior samples that are going to be used for imputation. The options for this step are described below.

4. Implement the NIMC application for the selected posterior samples.

5. Choose $\boldsymbol{\pi}^*$ by setting the values with the corresponding sliders and verify the results.

The NIMC provides the user with an immediate visualization of the missing and observed data patterns. Using this visualization, the user can repeat the step 5 until finding scenarios that reasonably reflect his or her beliefs about the missing data. We illustrate this process with data from the Colombian Annual Manufacturing Survey in Section 2.3.

Ranking the components

For the main task of choosing $\boldsymbol{\pi}^*$, it is helpful to have some method to identify the clusters. This can help to distinguish the clusters between MCMC iterations and organize the components for allocating new probabilities. We propose ranking the components based on a criterion related to $\boldsymbol{\mu}$ and $\boldsymbol{\pi}$, as we now describe.

In some situations where MNAR is considered, analysts may believe that the missing data have more extreme response values than the observed data. To represent this scenario, analysts can adjust the probabilities of the clusters that are located on the tails of the distribution. To facilitate such imputation modeling, we sort the components based on their distance δ to a reference point. In one-dimensional case, we would rank the mixture components based purely on the numerical order of μ. Extending that to the multivariate case, we can order the components based on their distance to the origin, $\delta_{\text{orig}} = \boldsymbol{\mu}'\boldsymbol{\mu}$, since the variables are standardized. To restrict the problem to vary in one direction, we suggest ranking them based on their distance to the minimum value on each dimension denoted by $\boldsymbol{y}_{\min} = (\min_i(\boldsymbol{y}_1), \ldots, \min_i(\boldsymbol{y}_p))$, where $\min_i(\boldsymbol{y}_j) = \min(y_{1j}, \ldots, y_{nj})$ for all $j = 1, \ldots, p$. The distance is denoted by $\delta_{\min} = (\boldsymbol{\mu} - \boldsymbol{y}_{\min})'(\boldsymbol{\mu} - \boldsymbol{y}_{\min})$. This makes the ranking more intuitive and interpretable, varying from the top cluster on the upper tail to the bottom cluster on the lower tail.

We note that any criterion other than δ_{orig} and $\delta_{\min}$ can be applied to this ranking stage. The ranking measure can be modified to be more suitable for each particular

imputation scenario. For example, if the data are negatively correlated, it might be more interpretable to consider the distance to a point other than the origin or the minimum data point.

After sampling from the MCMC, we order the posterior samples of the parameters based on $\delta_{\min}$. This post-simulation ordering is applied only to the occupied clusters; the empty components are not considered and are placed at the bottom of the list. Any clusters that are not representative of the nonrespondents can receive probability zero on the imputation step.

Ranking the components also helps with the problem of label switching that affects Bayesian mixture models. Because the model in (2.4) is invariant to permutations of the components, the posterior samples are not identifiable. To deal with this problem, methods that perform deterministic and probabilistic relabeling have been proposed (Rodríguez and Walker, 2014). For our purpose, an extensive method to deal with label switching is not necessary, since we do not intend to make component-specific inference. As long as the components are ranked based on some criterion, we can proceed to apply our methods to determine $\boldsymbol{\pi}^*$.

Selecting posterior samples

To generate completed data sets following the multiple imputation approach suggested by Rubin (1987), it is necessary to select m sufficiently spaced parameter samples from the MCMC iterations. Due to the mixture nature of the model (2.2)–(2.9), this subset of m posterior samples can include different cluster allocations. This requires specifying new probabilities $\boldsymbol{\pi}^*$ for every different allocation.

For simplicity, we propose summarizing the posterior cluster allocation by selecting only one iteration from the MCMC. Although technically this underestimates variability, we believe that analysts will find it easier to select $\boldsymbol{\pi}^*$ to generate data reflecting a particular set of beliefs. Having to select $\boldsymbol{\pi}^*$ multiple times for each

sensitivity analysis seems cumbersome, particularly since we do not know if any imputation model is "correct".

We select the sample that has the largest posterior value, similarly to the maximum a posteriori (MAP) discussed for normal mixture models by Fraley and Raftery (2007). After obtaining samples for the parameters from the Gibbs sampler after convergence, we evaluate the posterior at each iteration and select the one with the maximum posterior density value. With only one sample to summarize the posterior cluster allocation, the analyst can proceed to rank the components and change the probabilities as described in the previous sections. Then, the analyst can generate multiple imputed data sets from just the selected posterior sample and use the combining rules from Rubin (1987) for inference.

Implementation Issues

In this section, we discuss some implementation issues of using the proposed method to impute data with nonignorable missingness. One of the main parameters to be specified is the maximum number of mixture component K for the truncated Dirichlet process model. This value has to be sufficiently large to approximate well the distribution of the observed data and not be computationally impractical (Kim et al., 2014). As we mentioned before, we recommend starting with a maximum number of components between 30 and 50 and adjust based on the MCMC results. If the number of occupied clusters $\tilde{K}$ is too close to K, then the model may need to be refit with more components. However, a small value of $\tilde{K}$ is not necessarily the ideal, since it might limit the imputation alternatives as we noted in Section 2.2.1.

In both covariance specifications, the number of occupied clusters should be reasonable to provide flexibility for the imputation steps without becoming too large and overwhelming to specify $\boldsymbol{\pi}^*$. Based on our experience, a mixture with 10-15 occupied components seems to be an appropriate range for our purposes. The analyst

has to choose $\tilde{K}$ new probabilities for the imputation step. As this number increases, the task of choosing $\boldsymbol{\pi}^*$ becomes more cumbersome. To deal with this problem, the components can be combined into groups to reduce the dimension of the parameters that need to be specified. Then, the probabilities can be set jointly for all the components on each group. For example, the components in regions where no data should be generated can be combined into one group and receive probability zero. The user can also set the overall probability of a group and spread the value uniformly over its components, or leave the values based on the estimated probabilities from the observed data. The option to set the probabilities with the sliders can also be used to some clusters and be combined with any of the previous approaches to facilitate this step.

Even with a different distribution obtained by changing the probabilities, the imputation model is based on the clusters fitted to the observed data. If an analyst wants to define a scenario where some missing values exist where no data were observed, he can create new clusters. This requires specifying not only the new probabilities, but also the mean and covariance matrix of the new clusters. This requires more information about the missing data distribution. This information can be available in administrative records from other sources or previous surveys, and sampling restrictions related to the response variables. For example, on the U.S. Census of Manufactures, companies with less than five employees are not sent the forms. Their data can be imputed from administrative records from external sources, namely data from the Internal Revenue Service. In this case, a new cluster can be created based on the data available.

The implementation of the NIMC application facilitates sensitivity analysis, since it provides an automatic way for the user to compare the distributions of the observed and imputed data. Comparing distributions becomes more complicated, however, as the number of variables increases. There is no problem in using the proposed model

for large p if there is enough data to estimate more parameters. But, it may be not feasible to evaluate all pairwise scatterplots if there are more than say five to ten variables. In such cases, it can be useful to examine other measures, like values of key summary statistics of interest. The NIMC application provides means, quantiles, and correlations automatically as part of the output. The analyst can examine these statistics to see if the proposed $\boldsymbol{\pi}^*$ results in reasonable representations of the missing data with respect to these features of the distribution. When some variables are deemed more important for sensitivity analysis than others, we recommend that analysts focus their evaluations on the plots and summary statistics of those variables.

2.3 Illustrative Example

We illustrate the imputation procedure with data from the Colombian Annual Manufacturing Survey in 1991. The data comprise a total of 6609 plants and seven variables measured for each plant. The same data were used in Kim et al. (2014). To ease visualization of the method for this illustrative example, we focus on the results of the model with a subset of three response variables: RVA (real value added), RMU (real material used in products) and CAP (capital in real terms). In Section 2.3.1, we briefly illustrate the approach based on unequal covariances. Here, we only outline one imputation scenario without any formal sensitivity analysis, for reasons we explain later. In Section 2.3.2, we illustrate the approach based on equal, fixed covariances. Here, we include a sensitivity analysis with three imputation scenarios.

We also analyzed the results with all seven variables. The marginal and joint patterns of the observed data were very similar to what was observed with just three variables. After fitting the model with all variables, the number of clusters fitted and their location were also very similar to the results with the subset of variables.

Before proceeding to the model fitting, we use the data set with the variables RVA, RMU and CAP to create a nonignorable missingness pattern. To achieve that,

we simulate a missingness indicator $R_i \sim \text{Bern}(\theta_i)$ where the probability of missing θ_i is determined as $\theta_i = \text{logit}^{-1}(\beta_0 + \beta_1 \boldsymbol{y}_i)$ for $i = 1, \ldots, N$ with $\beta_0 = -2.3$ and $\beta_1 = 0.26$. We specify θ_i to reproduce the pattern in which individuals with larger quantities are more likely to not respond. The values of β_0 and β_1 are fixed such that the probability of missing is around 0.01 when $\boldsymbol{y}$ is around -3, and 0.5 when $\boldsymbol{y}$ is around 3, considering standardized variables. The data are heavily skewed, so we use the log transformation to visualize the variable relationships. The complete log transformed and standardized data set can be seen in the pairwise scatterplots in Figure 2.1. The mixture model is fitted to a total of $n = 5893$ observed points, plotted in gray.

2.3.1 Results with unequal covariances

First, we fit the model to the observed points with the vague prior specification described in Section 2.2.1. We set the maximum number of clusters to $K = 30$. To facilitate the application of the approaches for specifying $\boldsymbol{\pi}^*$, we select the MCMC iteration with the largest posterior value, as described on Section 2.2.2. The occupied top ranked clusters from the MAP iteration are summarized in Table 2.1 and plotted in Figure 2.2. We can see that most of the clusters with smaller probabilities have larger covariances and are more spread out than the clusters with higher probabilities, plotted with the darkest colors. Even though there are 11 occupied clusters, most of the weight (80%) is concentrated in four clusters.

Imputation scenarios With the clusters summarized in Table 2.1 and Figure 2.2, we can generate imputed data according to the guideline from Section 2.2.2 and set the new mixture probabilities. The values of $\boldsymbol{\pi}^*$ should reflect the expected proportion of nonrespondents on each cluster and result in a pattern of the imputed data in accordance with the analyst's beliefs. For example, if the analyst thinks that the

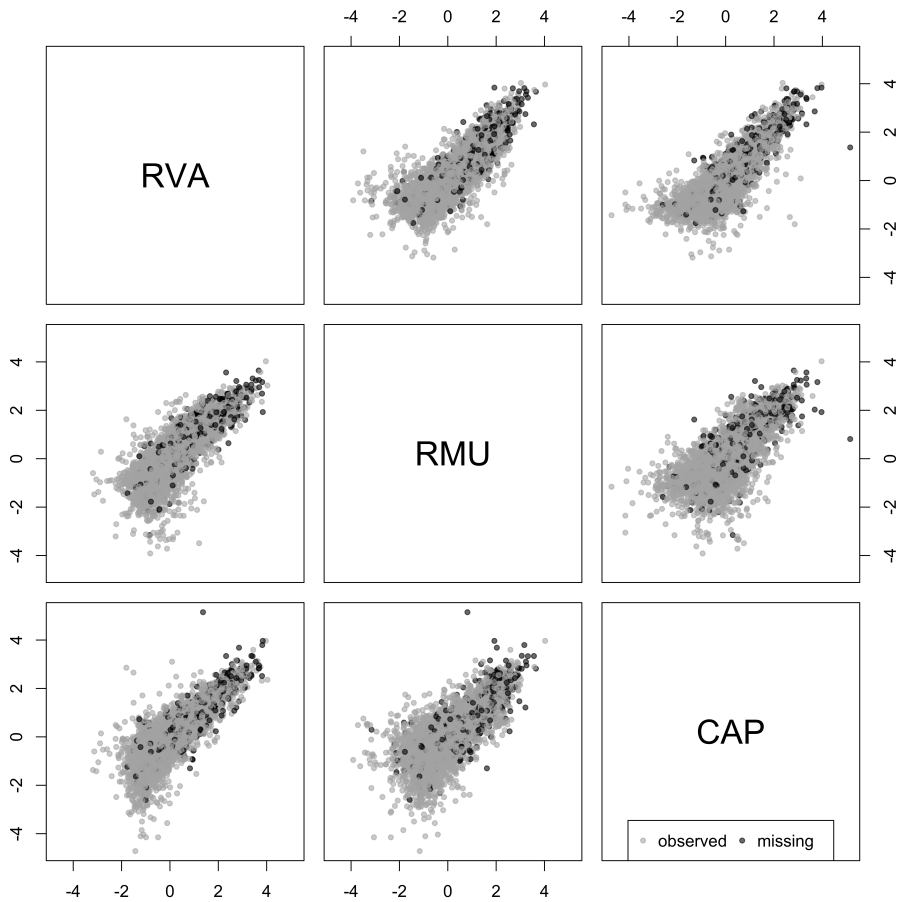


FIGURE 2.1: Complete Colombia data set (log transformed and positively standardized), with the 719 generated missing data points plotted in black and the 5893 observed data points plotted in gray.

nonrespondents tend to have larger values of $\boldsymbol{y}$, he can choose to assign all the weight to the top cluster. We can see a data set with missing data generated from this scenario in Figure 2.3. The other probabilities can be adjusted until the analyst is satisfied with the overall pattern of the missing data.

As we mentioned in Section 2.2.1, the default prior can yield spread and overlapping clusters such as in Figure 2.2. This can limit the possibilities of obtaining different targeted data patterns by tuning $\boldsymbol{\pi}^*$, since we cannot control the probabilities of smaller regions. For example, even with the probability allocated to just one cluster, like in Figure 2.3, the imputed data are spread over a large portion of the range of the observations because of the shape of the fitted clusters. If we want to

Table 2.1: Summary of the top ranked clusters on the MAP iteration for the Colombia data with unequal variances

cluster	μ_1	μ_2	μ_3	$\delta_{\min}$	π
1	0.892	0.843	0.843	70.175	0.242
2	0.457	-0.100	0.633	56.438	0.037
3	0.183	0.513	0.189	55.018	0.067
4	-0.166	-0.126	-0.192	43.951	0.280
5	-0.066	0.116	-0.719	41.956	0.024
6	-0.267	-0.684	-0.134	39.967	0.043
7	-0.784	-0.207	-0.350	38.588	0.021
8	-0.741	-1.371	0.280	37.426	0.004
9	-1.109	-0.091	-0.793	34.336	0.005
10	-0.727	-0.795	-0.485	33.694	0.148
11	-0.953	-0.804	-1.080	27.893	0.129

have more control over the imputation model and fine-tune the probabilities, we can use the model with fixed covariances suggested in Section 2.2.1, as we now describe.

2.3.2 Results with fixed covariances

It is harder to get imputed data localized in more specific regions with overlapping and spread clusters like in Figure 2.2. We now illustrate fixing the covariance matrix of all the components to force a larger number of separated components. The results are obtained using the same data displayed in Figure 2.1. With exception of the covariance matrix, the prior settings are the same as described in Section 2.3.1. We set the covariance matrices to be $\Sigma_k^{(t)} = \sigma I_p$, for all components $k = 1, \ldots, K$ and for all t. Since the variables are standardized, we recommend setting $\sigma < 1$ to result in smaller sized clusters. The choice of σ depends on the desired level of control for the imputation scenarios. The smaller the value of σ, the greater the control over small regions of the space. However, the number of occupied clusters will also increase, making the task of selecting $\boldsymbol{\pi}^*$ more complex as we discussed in Section 2.2.2.

In order to select an useful model for imputation purposes, we fit the model with different values of σ. We can see the summary of the MAP iteration from the results

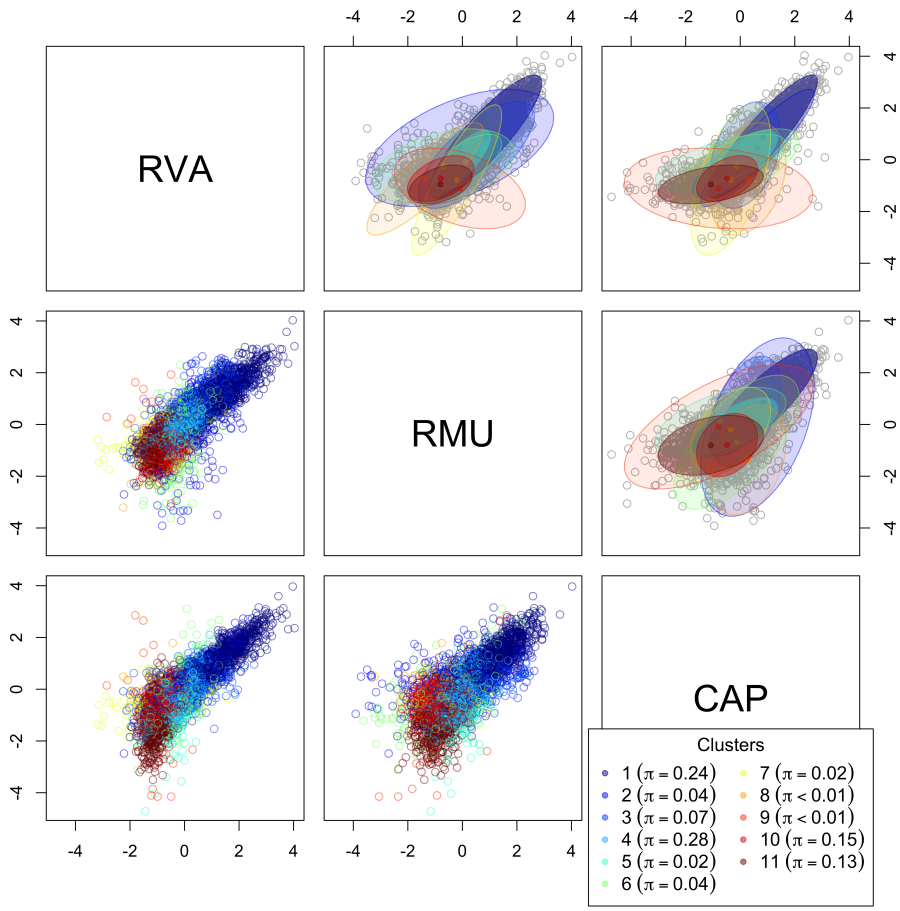


FIGURE 2.2: Summary of the top ranked occupied clusters on the MAP iteration for the Colombia data with unequal covariances. The points are colored by their cluster allocation on the lower diagonal, while the 95% quantile ellipses of the fitted clusters are plotted on the upper diagonal with color transparency proportional to the values of $\boldsymbol{\pi}$.

with $\sigma = \{0.1, 0.3, 0.5\}$ in Figure 2.4. With $\sigma = 0.1$, all the clusters are occupied, even though most have small probabilities. The cluster allocation in Figure 2.4(a) gives the analyst the option of imputing data in very small regions. Even though there are still many overlapping clusters in the center of the region, there are more clusters fitted in the tails with this value of σ. This can be useful in many MNAR applications, since it gives more control for generating extreme points.

However, the number of clusters fitted with $\sigma = 0.1$ is large making it somewhat difficult to choose $\boldsymbol{\pi}^*$. Some alternatives for facilitating this task are proposed in Section 2.2.2, like setting the combined probability for a group of clusters and keeping

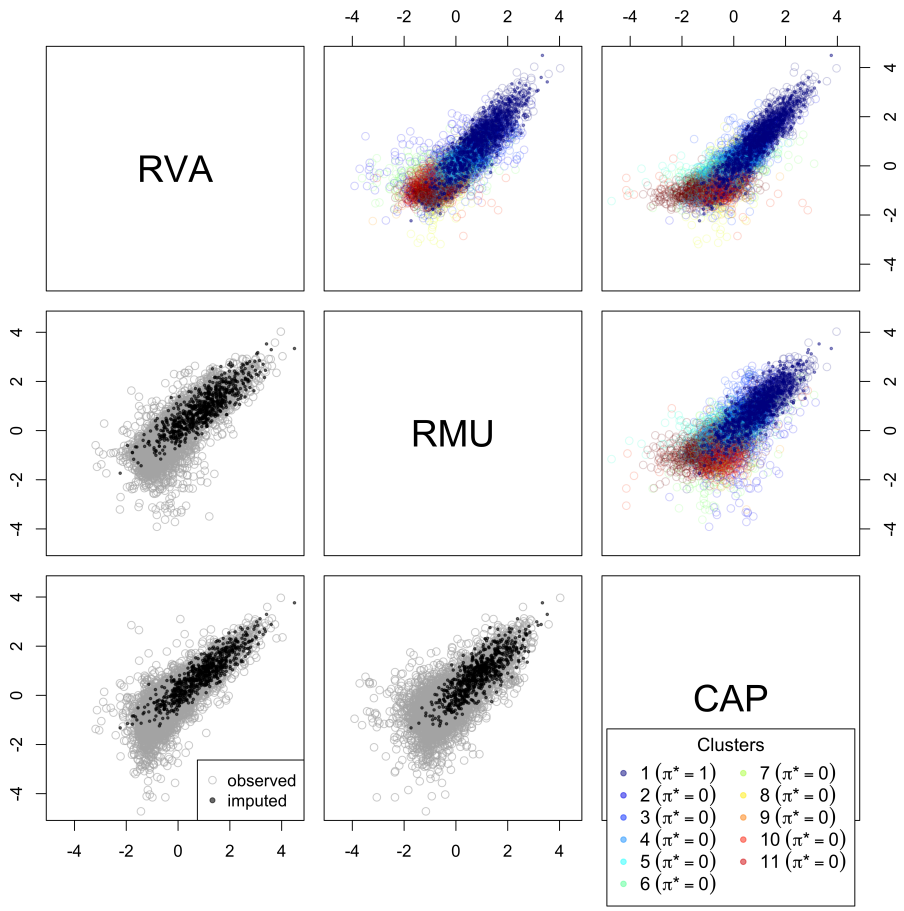


FIGURE 2.3: Complete Colombia data set generated from the model with unequal covariances and for imputation scenario: top cluster only. Observed points are plotted as hollow circles and imputed points are plotted as filled circles. The points on the lower diagonal are colored gray or black if they were observed or imputed, while the points on the upper diagonal are colored by the new cluster allocation.

some probabilities based on the estimated values. If that level of tuning is not necessary, the analyst can choose other values of σ. The model with $\sigma = 0.3$ seen in Figure 2.4(b) still allows for more control in the tails, but with a moderate number of clusters. With $\sigma = 0.5$, we obtained the same number of occupied clusters as seen in Figure 2.4(c). This variance, however, results in substantial mass in regions that we do not want to include in our imputations. Therefore, we select the model with $\sigma = 0.3$ to proceed with the imputation and sensitivity analysis.

Imputation scenarios After choosing the covariance value, we can proceed to the next step of selecting the new mixture probabilities. This is done with the NIMC application. The occupied top ranked clusters are summarized in Table 2.2, with the estimated probabilities from the observed data. As with the results with the unequal variances model, most of the clusters have small probabilities and 84% of the weight is concentrated in four clusters.

Some screenshots of the NIMC application with the results of data imputed with these estimated probabilities, i.e. under MAR, can be seen in Figure 2.5. The plots with the log transformed data, with and without standardization, are in Figure 2.5(a), while the plot in Figure 2.5(b) includes the data on its original scale without the log transformation. In this case, the skewness of the data distribution makes it hard to visualize the location of the fitted clusters. With the options available on the NIMC application, the user can compare both plots on the original scale to the plot of the standardized data used to fit the model. The choice of the original scale also affects the summary tables included in the second tab of the NIMC, as seen in Figure 2.6(a) on the log scale, and Figure 2.6(b) on the raw scale. Since these results are under MAR, the summary statistics of the complete imputed data and the original data are similar in all the different scales. With the values from the data on the raw scale in Figure 2.6(b), the analyst can analyze the imputation results on the same scale the data was recorded. In Figure 2.7, we can see a screenshot of the last tab with the generated complete data on the raw scale to be downloaded.

We present now the results of the imputation with probabilities set with the sliders of the NIMC application. Since the missing data plotted in Figure 2.1 were generated with larger probability of missing for larger values of $\boldsymbol{y}$, a possible imputation scenario is given by values of $\boldsymbol{\pi}^*$ decreasing from the top ranked cluster. A screenshot of the scatterplots with data generated under this scenario can be seen in Figure 2.8. We can see the probabilities set with the sliders on the left panel. Since we are dealing

Table 2.2: Summary of the top ranked clusters on the MAP iteration with fixed covariance matrices ($\sigma = 0.3$)

cluster	μ_1	μ_2	μ_3	$\delta_{\min}$	π
1	2.552	2.167	2.230	118.159	0.022
2	1.572	1.442	1.418	88.965	0.078
3	0.697	0.693	0.696	65.600	0.149
4	-0.951	0.815	1.456	65.477	0.000
5	0.888	-1.926	1.118	54.601	0.003
6	0.084	0.099	0.120	50.191	0.193
7	-0.057	0.082	0.121	49.167	0.012
8	-0.294	-0.093	-0.117	44.136	0.014
9	-0.583	-0.513	-0.459	36.477	0.350
10	-0.827	-0.787	-0.161	36.109	0.001
11	-0.970	-0.982	-1.050	26.960	0.147
12	-0.689	-2.635	-0.417	26.370	0.006
13	-1.166	-1.094	-2.390	17.441	0.024

with probabilities, the values set with the sliders have to be renormed to be used in the mixture model. The renormed probabilities are shown on the right of each slider on the panel on the left in Figure 2.8. The summary statistics are shown in Figure 2.9, for the standardized data on the left column and for the data on the raw scale on the right column. Unlike the MAR scenario, the values of the quantile statistics of both data scales increased when considering the imputed data with this pattern. When looking at the correlations of the standardized data, there is not much difference from the values observed under MAR because of the direction of the relationships between the variables. If the data are transformed back to the raw scale, we can see a larger difference on the correlations on the right column in Figure 2.9.

For comparison, we also include the results with data imputed from the top cluster only, that is, $\pi_1 = 1$ and $\pi_k = 0$ for $k = 2, \ldots, K$. The scatterplots with data on the standardized scale and on the raw scale can be seen on the screenshot of the application in Figure 2.10, and the summary statistics on the screenshots in Figure 2.11. As with the scenario with the decreasing probabilities, there are some increases

in the quantiles of the complete data from concentrating the imputed values on high values of $\boldsymbol{y}$. Because we are generating data from such an extreme scenario, the summary statistics of the imputed data are very different from the combined and observed data sets. This impact on the summary statistics is especially large on the correlation matrices of the imputed data. Since all the points are sampled from a single normal distribution with fixed covariance $\Sigma = \sigma I_p$, the correlations are close to zero by construction. If we look at just the observed data classified at the top cluster, the correlations range from 0.3-0.5. We can note this difference in the pattern of the observed data and the area corresponding to the top cluster in the plots in Figure 2.4(b), where the observations plotted as the gray circles are not spread over the entire purple circle of the top cluster on the scatterplots in the upper diagonal. If the user wants to be more restrictive about the values the imputed data can assume, he can either decrease the size of the clusters or use the modified hot-deck imputation approach that we mention in Section 2.2.2.

The point of generating imputations in three (or more) scenarios is to evaluate the sensitivity of inferences to different assumptions about the missing data. To illustrate this, we perform a sensitivity analysis with the marginal means of each variable. To do so, for each $\boldsymbol{\pi}^*$ we generate $m = 5$ completed datasets, each containing the observed data and a realization of the NIMC imputation process. Following Rubin (1987), we compute the means and variances using the usual multiple imputation combining rules. The point estimates and confidence intervals for the three scenarios (MAR, decreasing probabilities and top cluster only) are included in Table 2.3. As expected, we can see that the impact of the MNAR assumptions on the inference for the marginal means, as the estimates increase as we move along the imputation scenarios. With these data, apparently the marginal means are highly sensitive to these assumptions about the missing data. If analysts really believe such imputation scenarios are plausible, they should endeavor to obtain additional information on

Table 2.3: Point estimates and 95% confidence intervals for the marginal means calculated with the combining rules for multiple imputation with Colombia data generated under different scenarios.

(a) Log transformed and standardized data

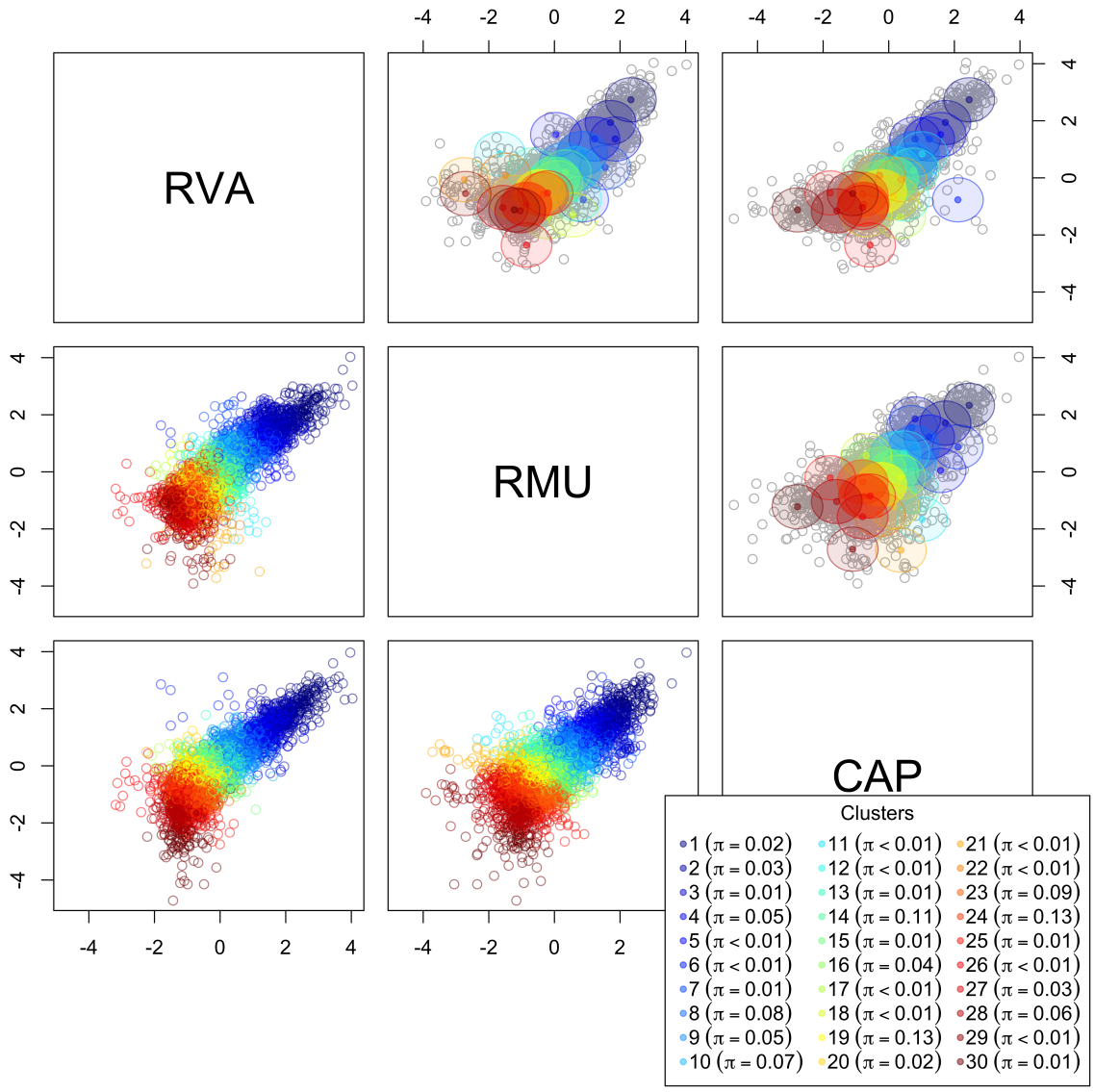
	MAR	Decreasing probabilities	Top cluster
RVA	-0.088 (-0.113,-0.063)	0.089 (0.057,0.122)	0.202 (0.17,0.233)
RMU	-0.081 (-0.11,-0.053)	0.08 (0.051,0.109)	0.164 (0.134,0.194)
CAP	-0.079 (-0.104,-0.053)	0.093 (0.064,0.122)	0.176 (0.146,0.206)

(b) Original scale data

	MAR	Decreasing probabilities	Top cluster
RVA	16650 (14511,18790)	33104 (28120,38087)	53351 (48173,58530)
RMU	21074 (17869,24279)	38941 (33232,44650)	57075 (51229,62922)
CAP	16399 (13292,19507)	35942 (30292,41592)	55153 (47215,63091)

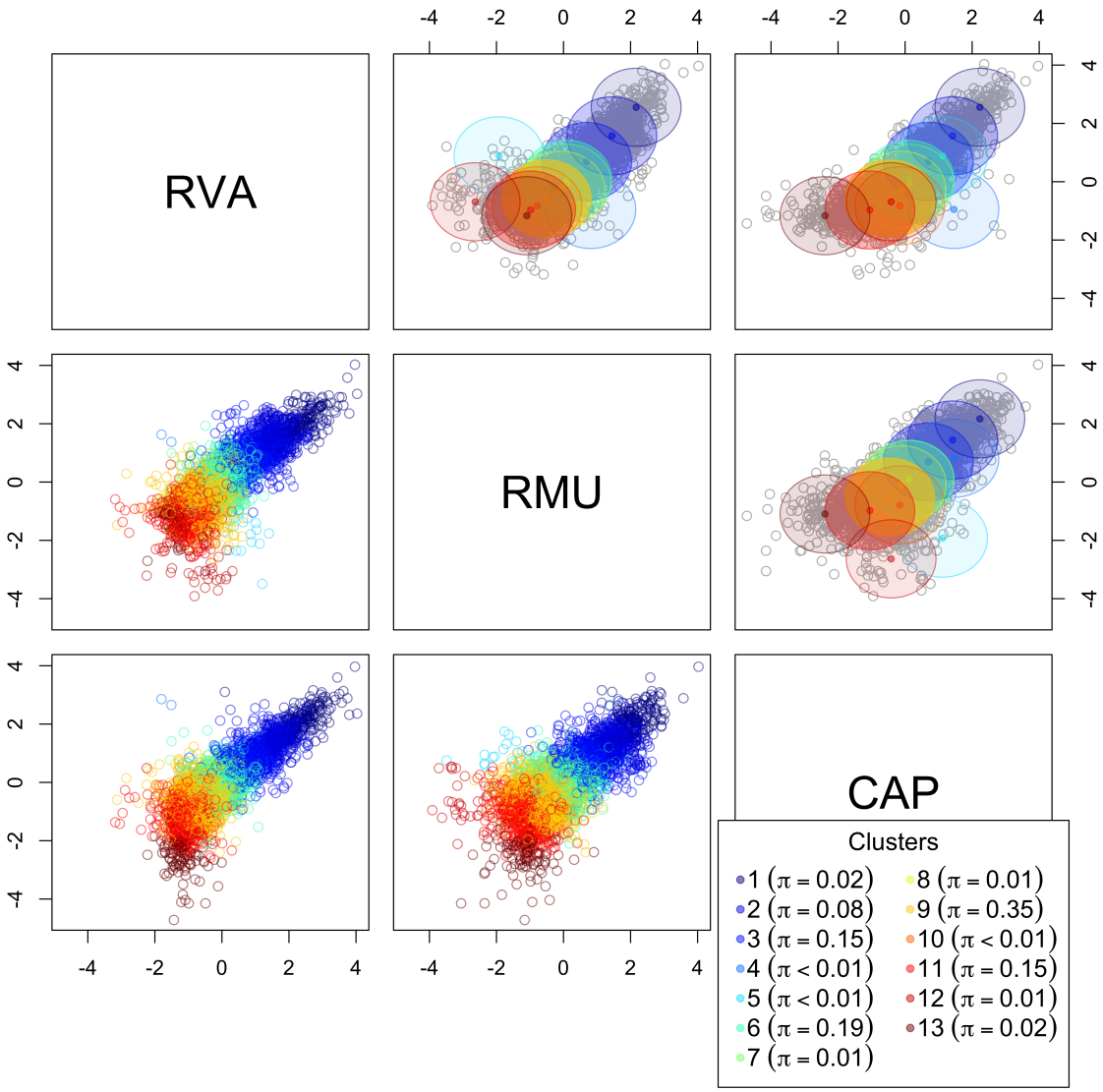
the missing values, for example via external data sources or nonresponse followup sampling. We discuss the latter strategy in Chapter 3.

The user interface of the NIMC application enables the visualization of any imputation scenario on the spot. With the scatterplots, the analyst can determine the missing data pattern for smaller regions of the data on different scales. Analysts also can compare the impacts of different scenarios on summary statistics and distributions. Such comparisons help analysts to assess the sensitivity of inferences to different assumptions about the missing data.



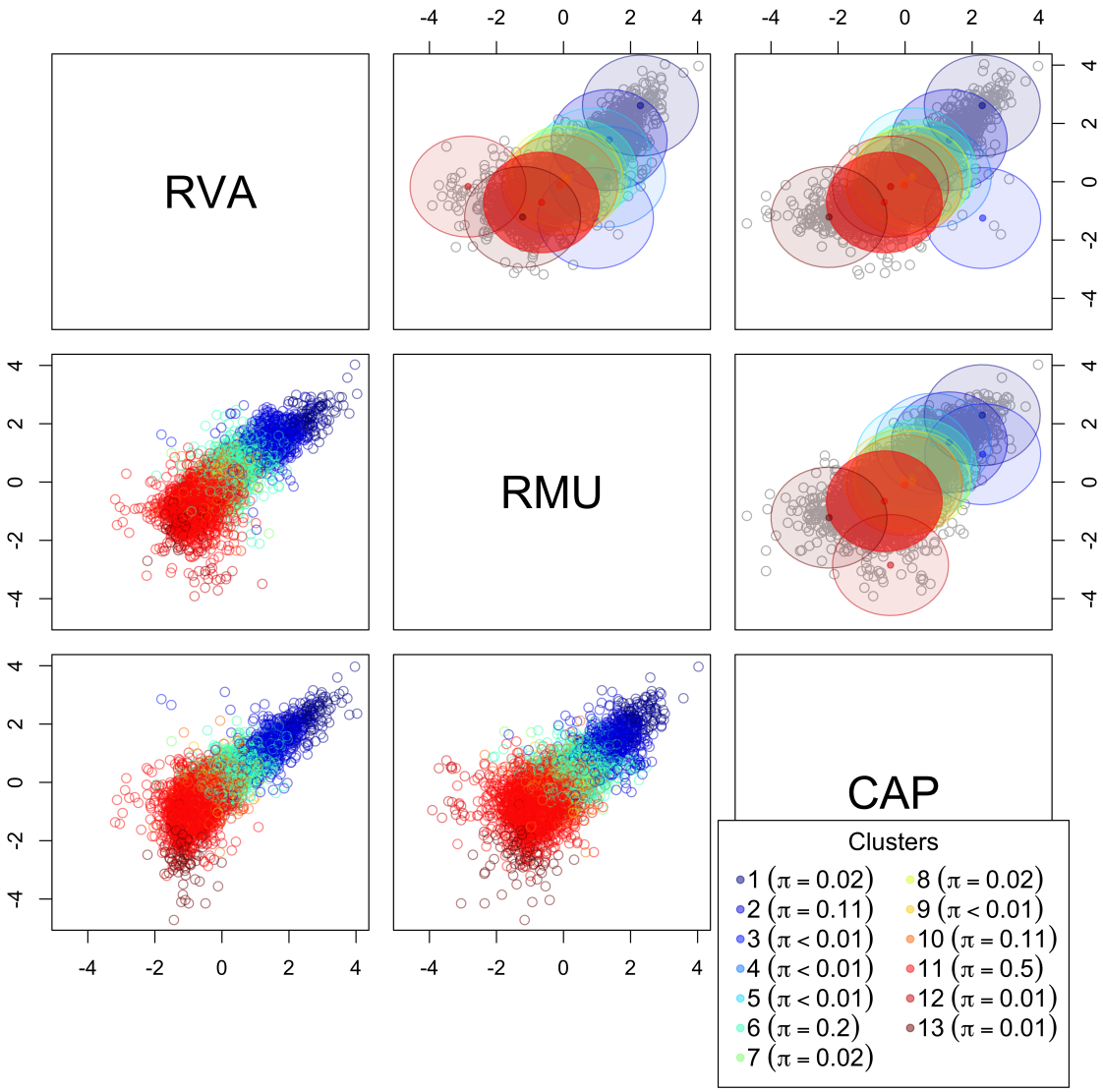
(a) Results of the model with fixed covariances and $\sigma = 0.1$

FIGURE 2.4: Summary of the top ranked occupied clusters on the MAP iterations for the Colombia data set with fixed covariances. The points are colored by their cluster allocation on the lower diagonal, while the 95% quantile ellipses of the fitted clusters are plotted on the upper diagonal.



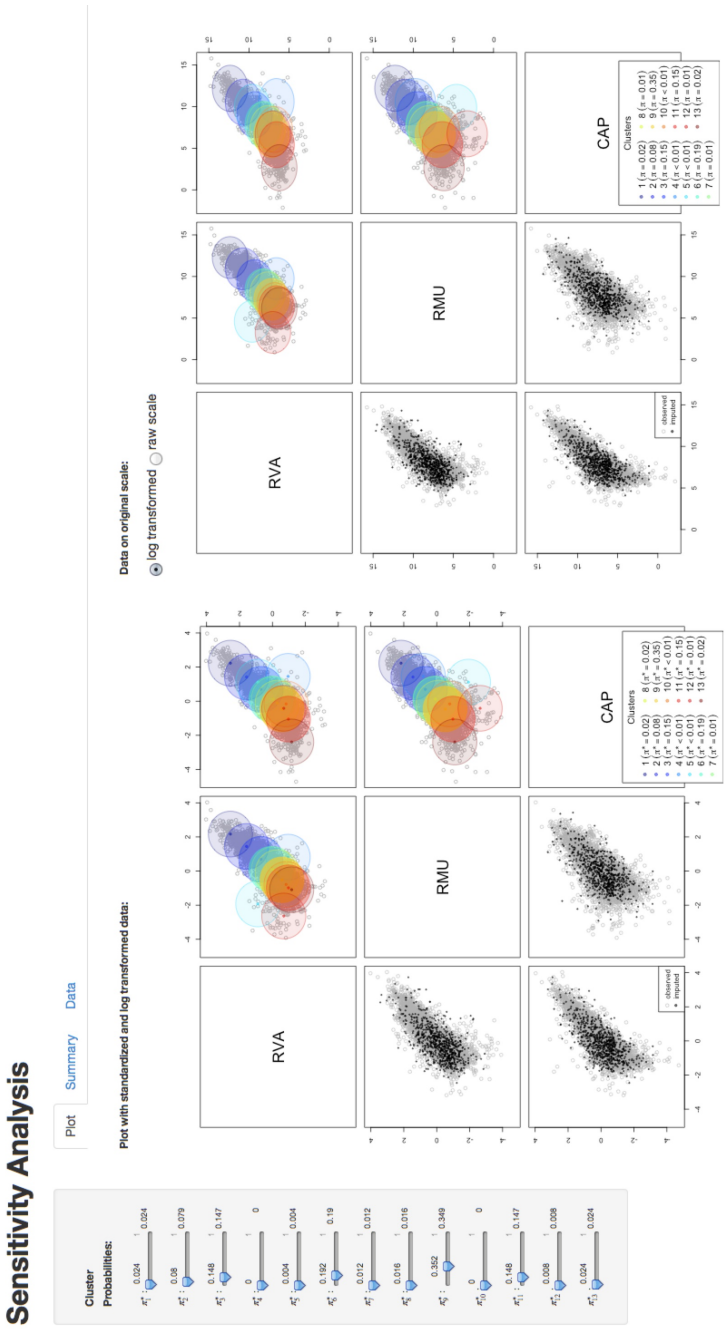
(b) Results of the model with fixed covariances and $\sigma = 0.3$

FIGURE 2.4: Summary of the top ranked occupied clusters on the MAP iterations for the Colombia data set with fixed covariances. The points are colored by their cluster allocation on the lower diagonal, while the 95% quantile ellipses of the fitted clusters are plotted on the upper diagonal.



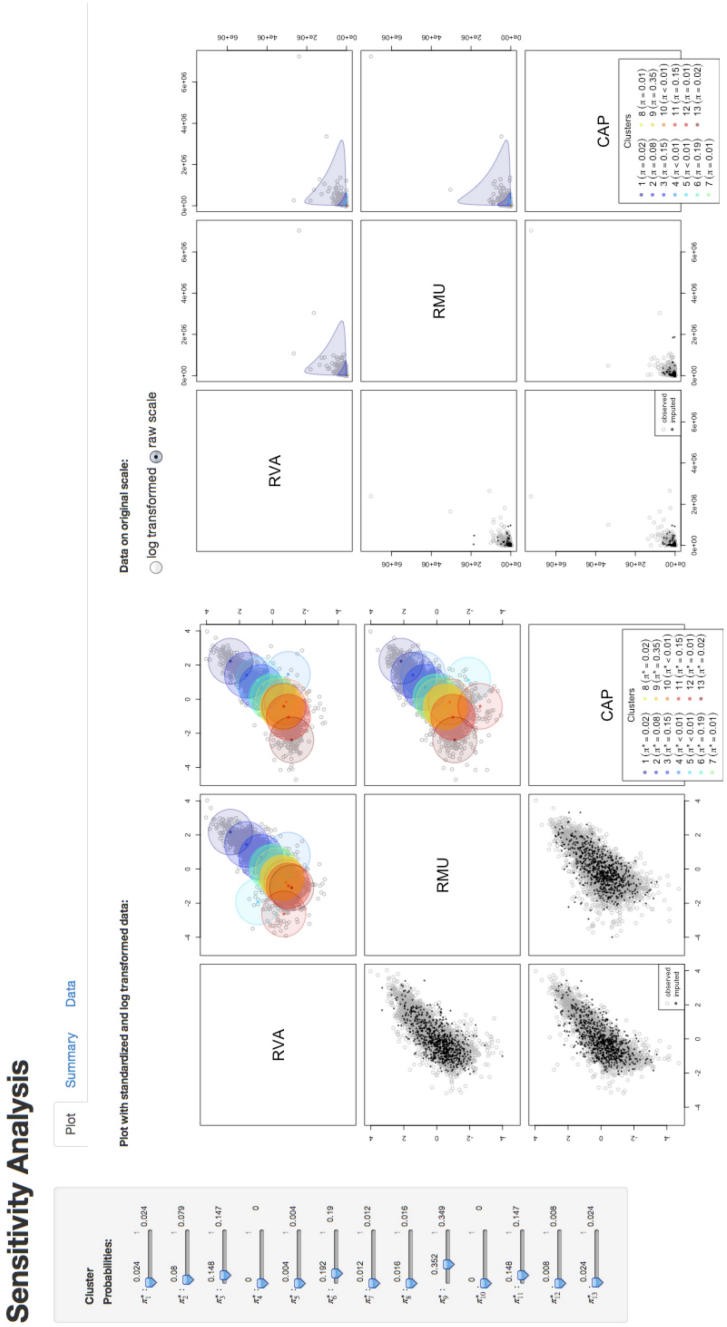
(c) Results of the model with fixed covariances and $\sigma = 0.5$

FIGURE 2.4: Summary of the top ranked occupied clusters on the MAP iterations for the Colombia data set with fixed covariances. The points are colored by their cluster allocation on the lower diagonal, while the 95% quantile ellipses of the fitted clusters are plotted on the upper diagonal.



(a) Pairwise scatterplot of the data on the log transformed and standardized scale on the left, and with the log transformation option selected on the right.

FIGURE 2.5: Screenshots of the plot tab of the NIMC application with the Colombia data and imputed data generated with the estimated values of $\boldsymbol{\pi}$, assuming MAR.



(b) Pairwise scatterplot of the data on the log transformed and standardized scale on the left, and with the raw scale option selected on the right.

FIGURE 2.5: Screenshots of the plot tab of the NIMC application with the Colombia data and imputed data generated with the estimated values of $\boldsymbol{\pi}$, assuming MAR.

(a) Summary statistics of the data on the log transformed and standardized scale on the left column, and with the log transformation option selected on the right column.

FIGURE 2.6: Screenshots of the summary tab of the NIMC application with the Colombia data and imputed data generated with the estimated values of $\boldsymbol{\pi}$, assuming MAR.

(b) Summary statistics of the data on the log transformed and standardized scale on the left column, and with the raw scale option selected on the right column.

FIGURE 2.6: Screenshots of the summary tab of the NIMC application with the Colombia data and imputed data generated with the estimated values of $\boldsymbol{\pi}$, assuming MAR.

FIGURE 2.7: Screenshots of the data tab of the NIMC application with the Colombia data and imputed data generated with the estimated values of $\boldsymbol{\pi}$, assuming MAR.

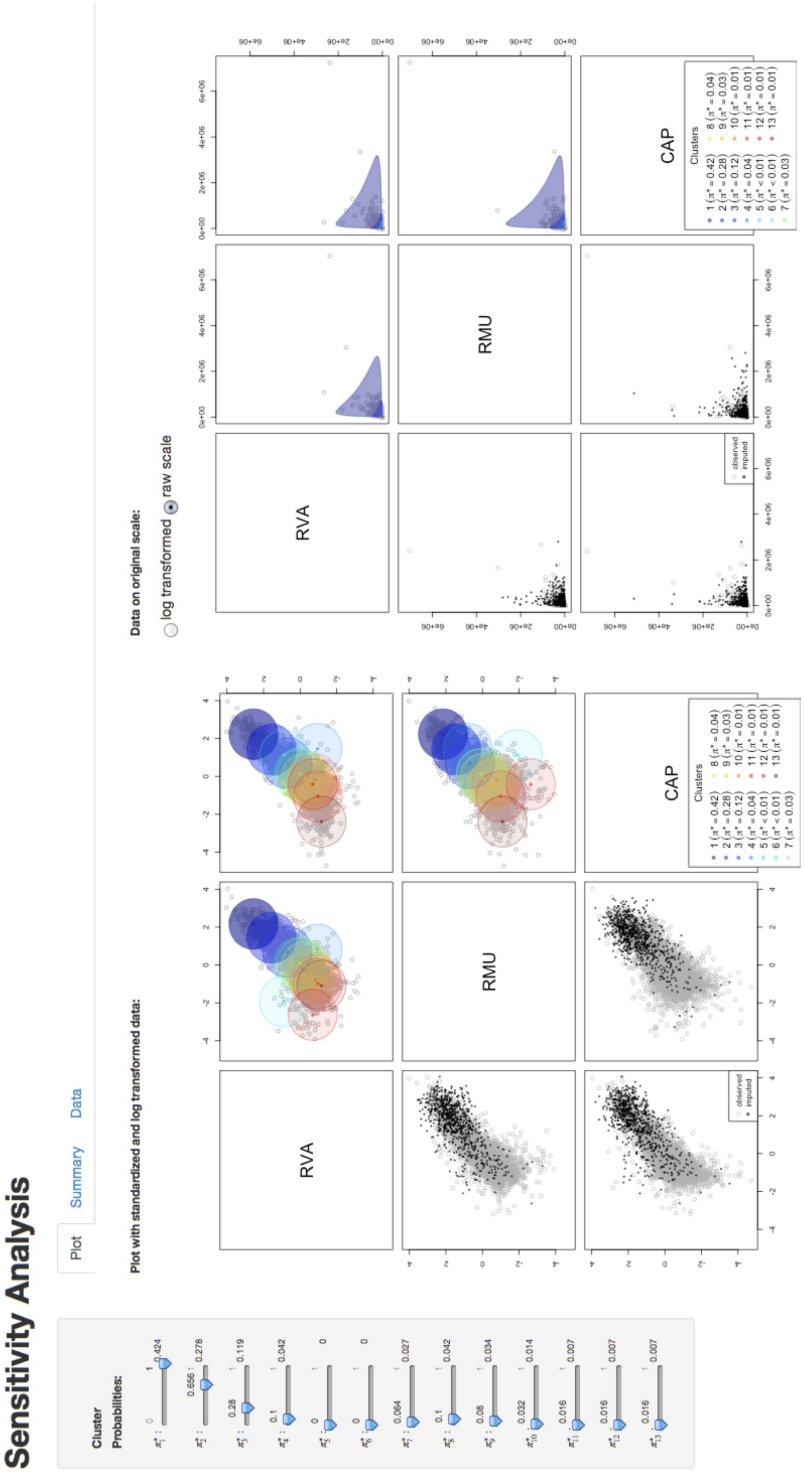


FIGURE 2.8: Screenshot of the plot tab of the NIMC application with the Colombia data and imputed data generated with decreasing probabilities $\boldsymbol{\pi}^*$. The scatterplot on the right includes data plotted on the original raw scale.

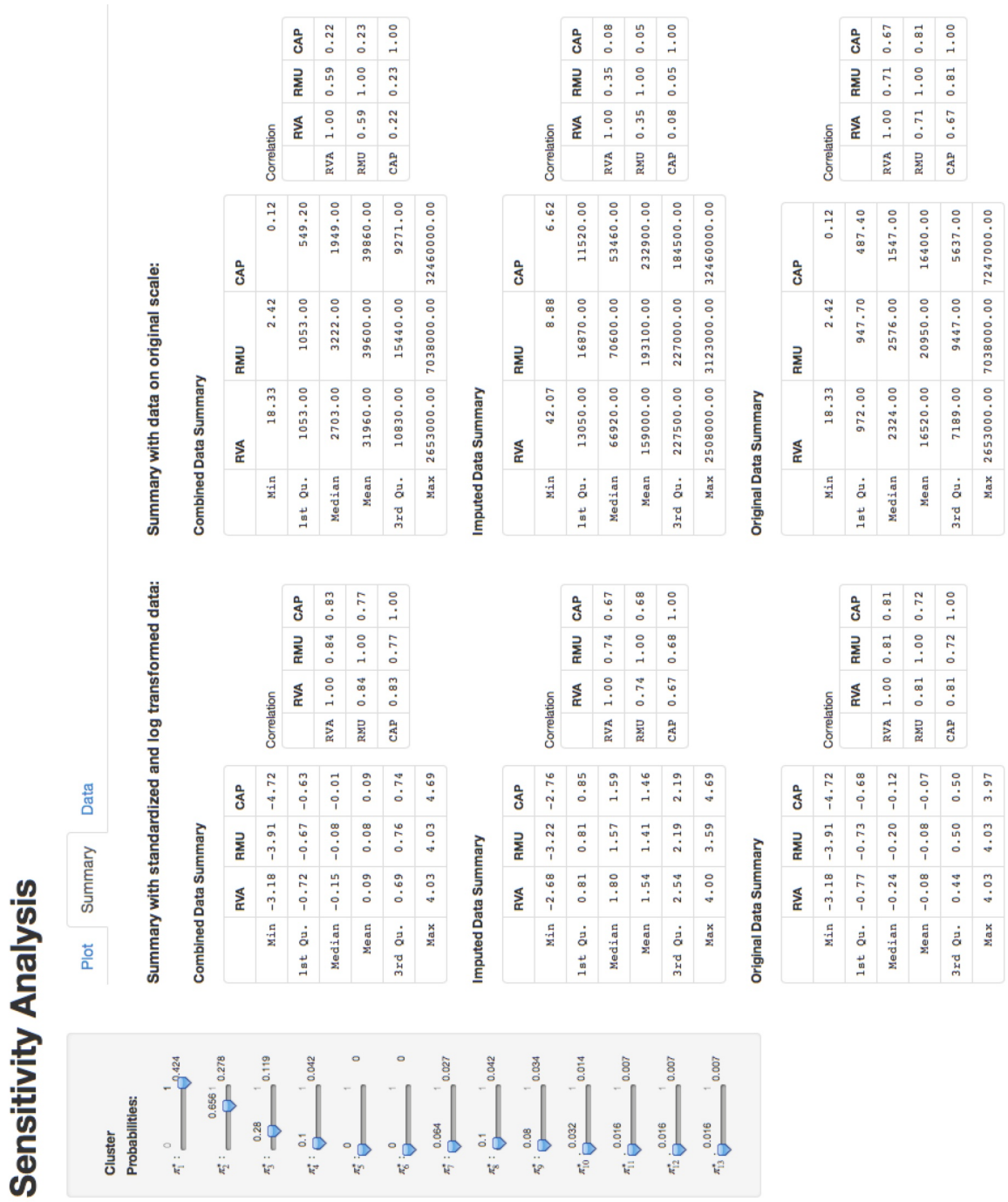


FIGURE 2.9: Screenshot of the summary tab of the NIMC application with the Colombia data and imputed data generated with decreasing probabilities $\boldsymbol{\pi}^*$. The values on the right column are calculated with data on the original raw scale.

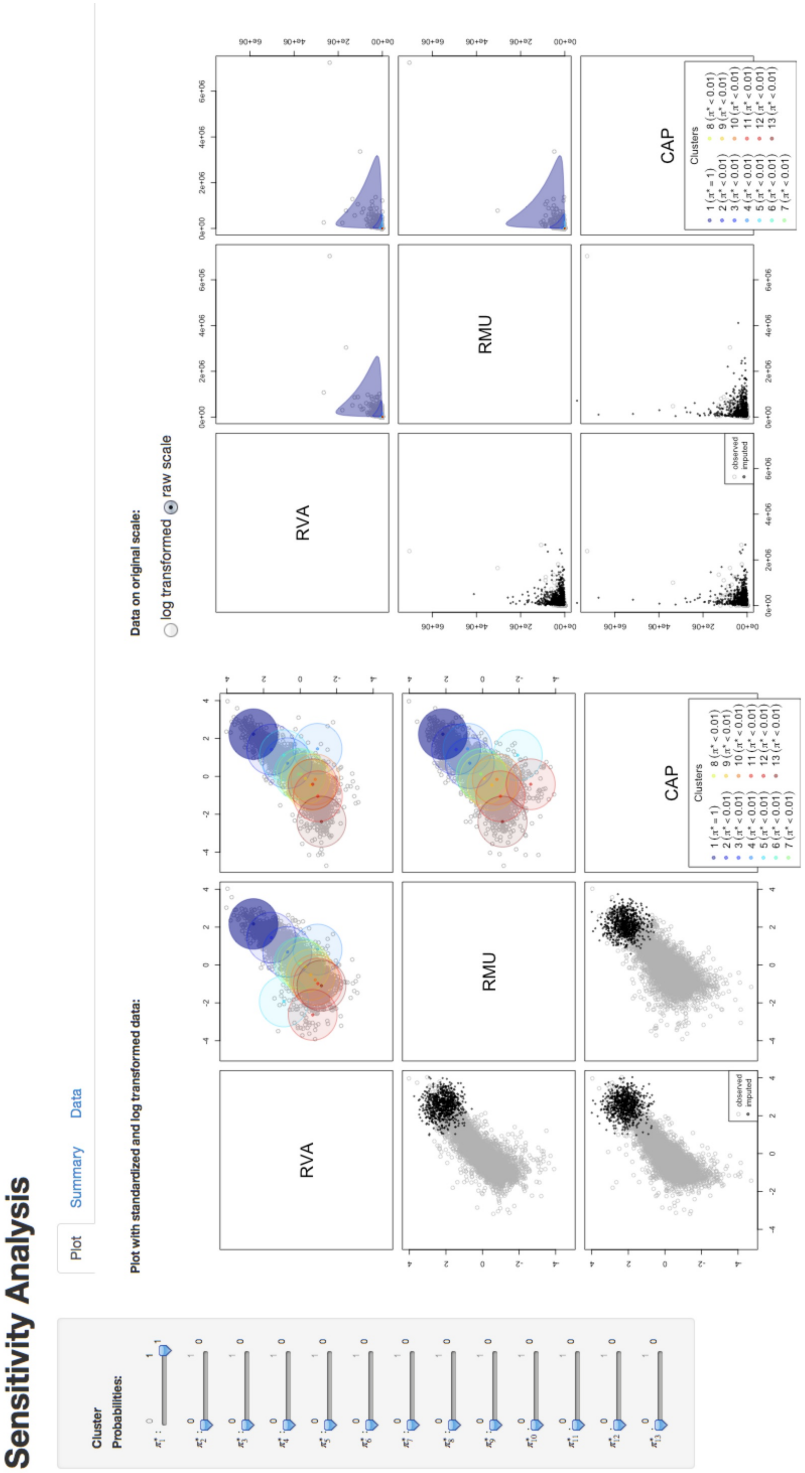


FIGURE 2.10: Screenshot of the plot tab of the NIMC application with the Colombia data and imputed data generated with all probability allocated to the top cluster only. The scatterplot on the right includes data plotted on the original raw scale.

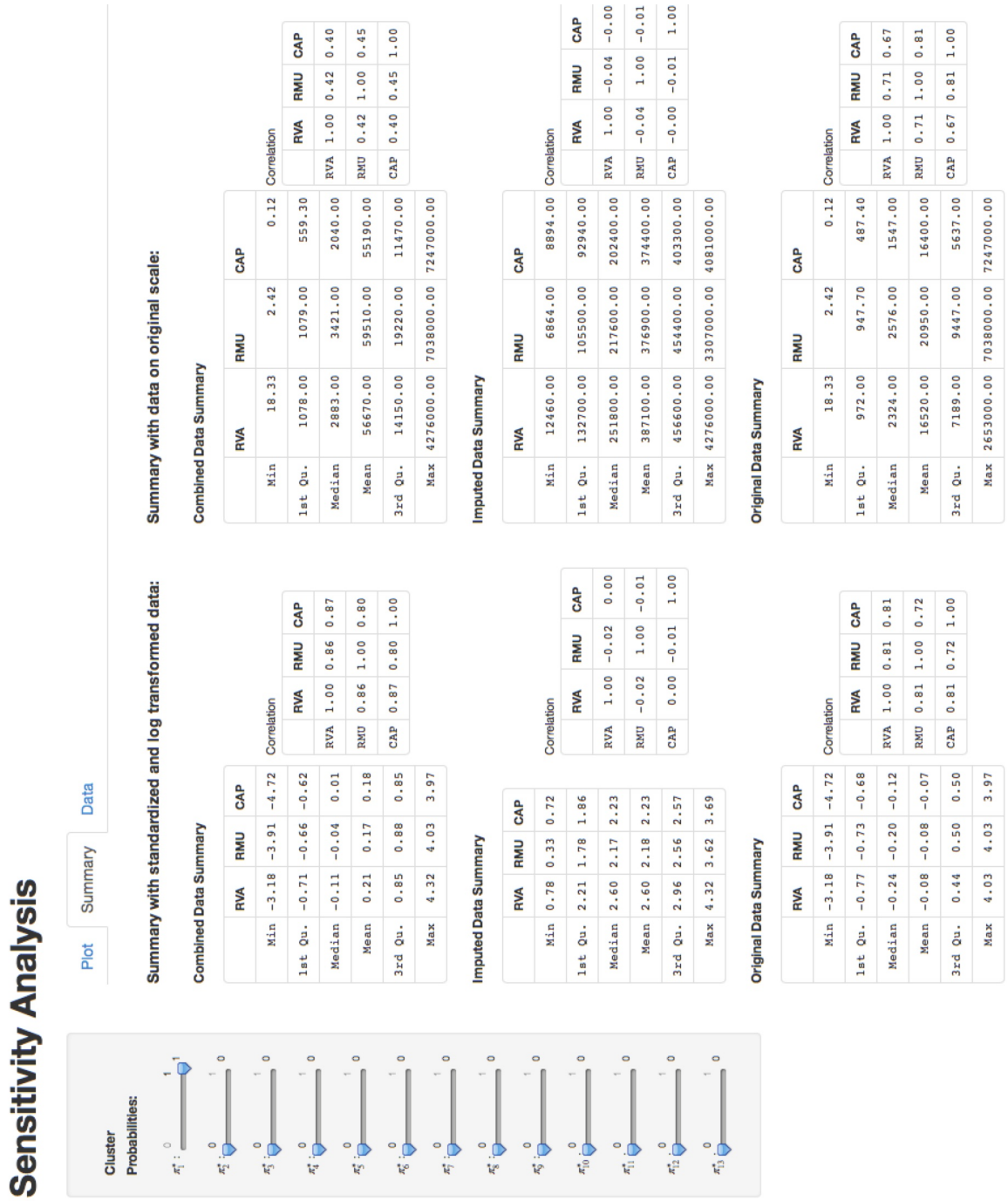


FIGURE 2.11: Screenshot of the summary tab of the NIMC application with the Colombia data and imputed data generated with all probability allocated to the top cluster only. The values on the right column are calculated with data on the original raw scale.

2.4 Simulations

In this section, we apply the methods described in Section 2.2 to repeated samples of a simple simulated data set. The goal of this example is to show that the true complete data distribution can be recovered if the missing data mechanism is known, so that the method can be interpreted as a means for sensitivity analysis.

For each of 500 repetitions, we generate a sample of size $n = 1000$ from a mixture of three two-dimensional normal distributions with parameters

$$\boldsymbol{\pi}_0 = \begin{bmatrix} 0.33 \\ 0.33 \\ 0.33 \end{bmatrix}, \qquad \boldsymbol{\mu}_0 = \begin{bmatrix} 0 & 0 \\ 4 & 5 \\ 10 & 6 \end{bmatrix}, \qquad \Sigma_{k0} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \text{ for } k = 1, 2, 3.$$

For each observation i, the probability of missing depends entirely on the cluster it belongs to, denoted by $z_i \in \{1, 2, 3\}$. Thus, we sample the indicator of missingness $R_i \sim \text{Bernoulli}(\theta_{z_i})$, where $\boldsymbol{\theta} = \begin{bmatrix} 0.1 & 0.1 & 0.4 \end{bmatrix}'$. Figure 2.12 displays one realization of the simulated data. The observations plotted in gray, after being standardized, are used for fitting the mixture model and simulating $m = 5$ imputed data sets for each repetition. To avoid fitting a different number of clusters than what was specified, we fix $K = 3$ and add a verification step at each Gibbs sampler iteration to restart if the number of occupied components is different than K.

With the missing data mechanism completely specified, we know that the posterior samples of each π_k should converge to the proportion of observations in each cluster. Similarly, we can determine the theoretical proportions of missing data per cluster, that is approximately $\pi_{k0} \times \theta_k$, and fix that as the new mixture weights. Since the posterior samples are reordered based on the distance to the origin as described in Section 2.2.2, the proportions are also reordered to match the clusters.

In order to evaluate the performance of the imputation process, we make inference about some quantities on each of the completed original data sets, and with just

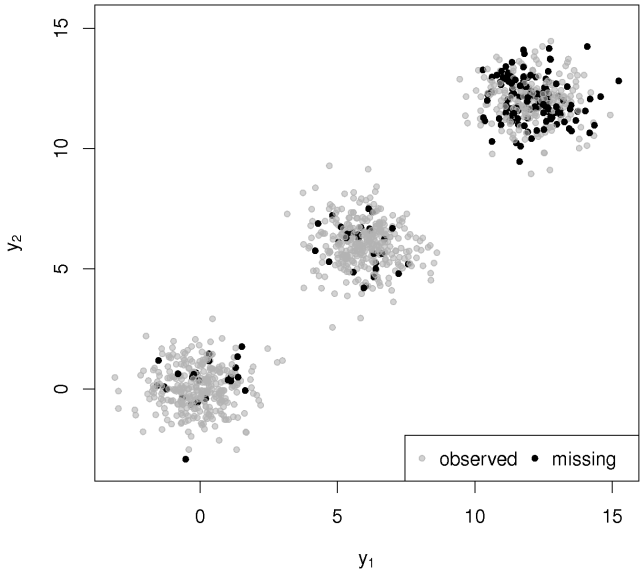


FIGURE 2.12: Example of one realization of a complete simulated data set for the simulated example, with the points plotted by the missing status

the observed data with no imputation. We also use the combining rules for multiple imputed data from Rubin (1987), reviewed in Chapter 1, to make inference on the imputed data sets using our method assuming MNAR, as well as comparing to imputation assuming MAR. The quantities analyzed are the marginal means $\bar{y}_1$ and $\bar{y}_2$, as well as the regression coefficients from the linear model of y_2 on y_1. In Figure 2.13, we can see the 95% confidence intervals of the marginal means for each repetition under the four different approaches. The results for the regression coefficients are plotted on Figure 2.14. For the marginal means, the truth, plotted as the dashed line, is obtained by the theoretical mean of the mixture model. For the regressions, the truth are the coefficients from the model fitted with all the 500 original complete data sets combined.

From the coverage plots and rates on Table 2.4, we can see how the marginal means are really underestimated when using just the observed data or when imputing data assuming missing at random. For the regression coefficients, the results are not

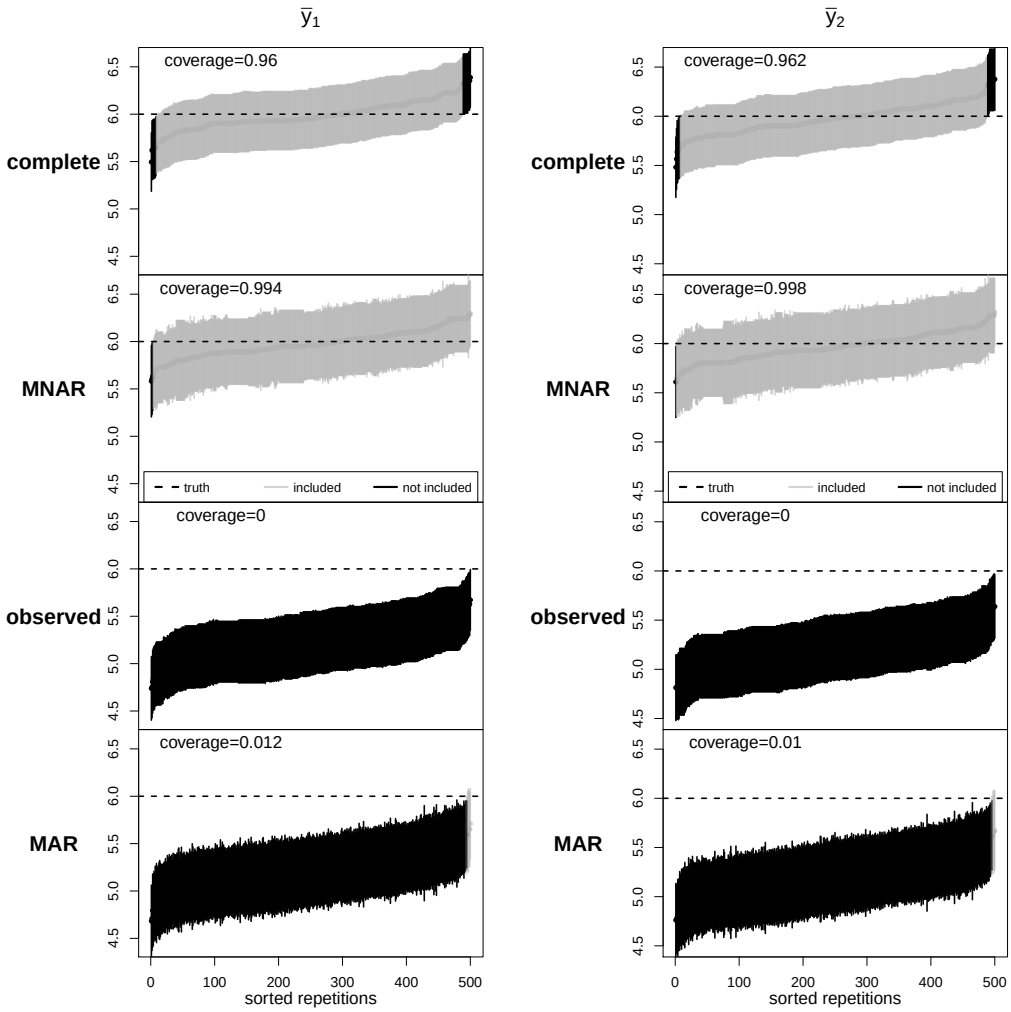


FIGURE 2.13: Coverage of confidence intervals of marginal means using: complete original data, imputed data assuming MNAR, just the observed original data, and imputed data assuming MAR. The intervals are sorted by their center points and are plotted in gray or black if they cover the truth (dashed line) or not.

affected as much due to the linear pattern of the data even with the nonignorable missingness. As expected, our method, denoted by MNAR, performs well for all estimated quantities with results similar to those obtained with the original complete data with no missing data.

Due to the clustered pattern of the data, looking at point estimates like the marginal means and the regression coefficients is not enough to guarantee that the imputation methods are generating data correctly. We could obtain means and

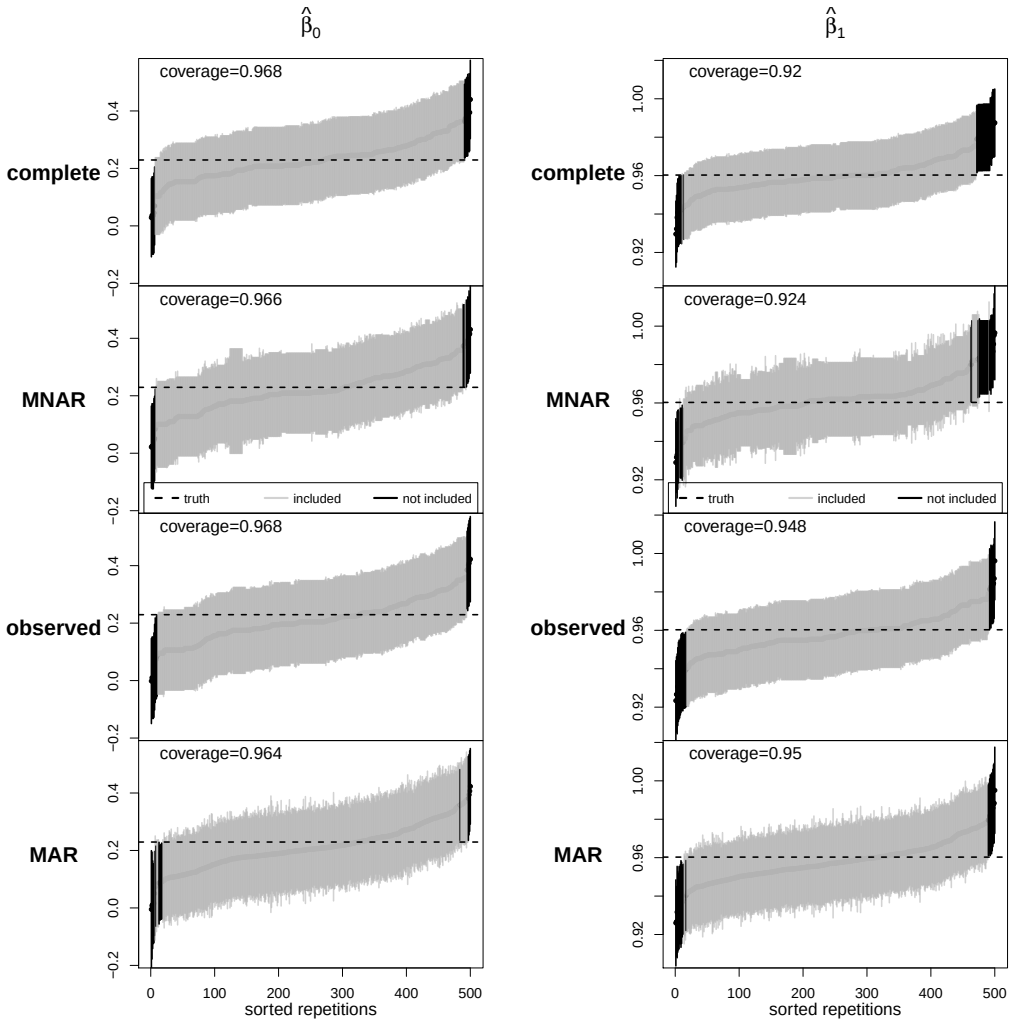
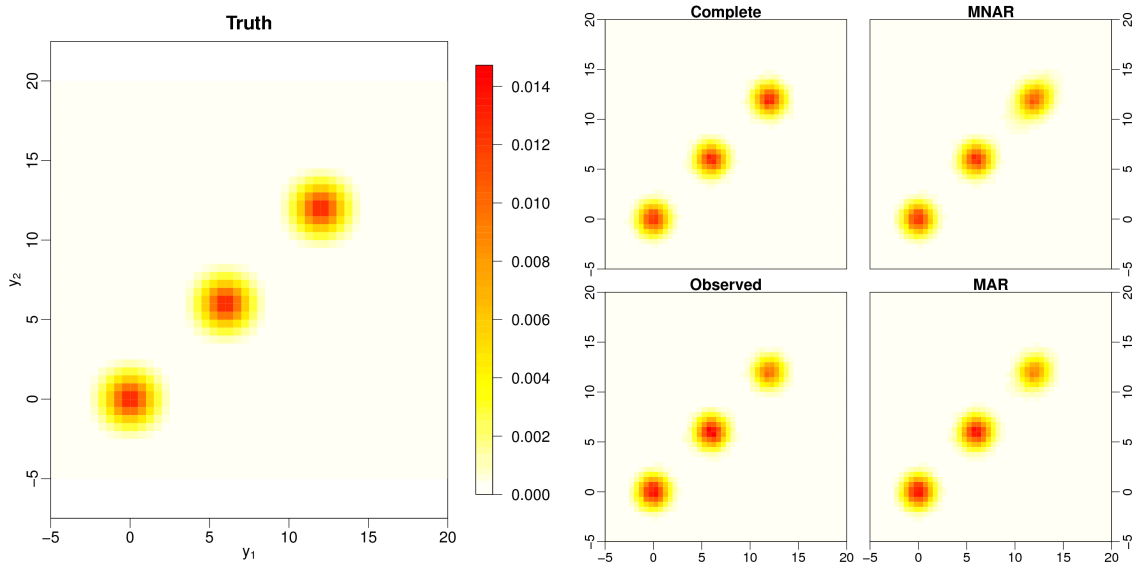


FIGURE 2.14: Coverage of confidence intervals of regression coefficients using: complete original data, imputed data assuming MNAR, just the observed original data, and imputed data assuming MAR. The intervals are sorted by their center points and are plotted in gray or black if they cover the truth (dashed line) or not.

coefficients similar to the truth with data from a pattern different from the target, for example if the points were placed into a large unified cluster with the appropriate location and correlation. For that reason, we also investigate the differences in the estimated frequency of points in an uniform grid over the range of observations. The truth in this case is an approximation to the normal density evaluated at the grid cells plotted in Figure 2.15(a). The average frequencies over the repeated samples for the complete, imputed under MNAR, observed and imputed under MAR data

Table 2.4: Coverage rates of the different estimates under the four approaches

	$\bar{y}_1$	$\bar{y}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$
complete	0.96	0.96	0.97	0.92
MNAR	0.99	1.00	0.97	0.92
observed	0.00	0.00	0.97	0.95
MAR	0.01	0.01	0.96	0.95



(a) (b)

FIGURE 2.15: Frequency maps of the true density on (a) and the estimated with the different data sets on (b)

sets are plotted in Figure 2.15(b) with the same color scale. The average frequencies with the complete data and with the MNAR imputation are more similar to the true density, whereas the frequencies with the observed data and with the MAR imputation underestimate the height of the top cluster.

For this simulation study, we used a simple data setup to enable us to control the true data generating mechanism. Of course in practice one does not know the true value of $\boldsymbol{\pi}^*$. As we illustrated in Section 2.3, the choice of the imputation probabilities depends on the analyst's prior beliefs and can be evaluated for different scenarios. The simulation results suggest that the method offers a meaningful way of doing sensitivity analysis.

2.5 Conclusion

We present a flexible approach to perform multiple imputation for multivariate continuous data with nonignorable missingness. We estimate the observed data distribution with a Dirichlet process mixture of multivariate normals. This is able to approximate well many data patterns, and can be easily sampled from with the conjugate prior specification. Moreover, the mixture model provides the user with the clustered observations and a distribution easy to modify. Under this framework, any type of prior information about the missing data can be incorporated into the model. For the scope of this paper, we considered only cases of unit nonresponse, so that the method generates the entire vector of responses. In case of item nonresponse, the missing variables can be generated within the MCMC by sampling from the conditional multivariate normal distributions given the cluster label. This assumes the item nonresponse is missing at random. We demonstrate this alternative in the next chapter.

Considering the missing not at random assumption imposes an extra level of difficulty, since the true complete data distribution is never known. Because of this limitation, it is important that an imputation method can be easily changed to reflect different hypothesis. Since our method is based on the distribution fitted to the observed data, after obtaining the posterior samples and using the NIMC application, it is straightforward to change the weights and generate new imputations. This is essential for a sensitivity analysis where the possibilities are compared until the target pattern is achieved.

The method is built based on the characteristics of the observed data distribution. While this can facilitate the imputation step, it can also limit the imputed data to the patterns that are already observed. For example, by basing the new probabilities on the estimated values we can be underestimating the true parameters of the complete

data. Another limitation is when there is reason to believe that the nonrespondents should be located in a different region than the respondents. In that case, our method can still be used by using this strong belief to specify a new cluster.

The results presented here provide some guidance into the possible ways to generate imputed data and illustrate how this can be done with some visual tools. This approach accommodates many scenarios while maintaining the interpretability of the task of translating the analyst's beliefs into the parameter space.

3

Using imputation techniques to evaluate stopping rules in adaptive survey designs

3.1 Introduction

Many survey organizations are considering adaptive survey design techniques (Miller, 2013; Finamore et al., 2013) to improve the sample representativeness while allocating resources more efficiently. Adaptive designs use auxiliary information to tailor and update the sampling scheme throughout the survey. This information may include administrative records; paradata, that is data about collecting the responses, such as details about where and how the interviews took place; and the actual survey data as they are collected. With these extra information available, the agency conducting the survey can modify features of the original sampling scheme. Some of these changes affect the strategies applied to individuals, e.g., using different contact means or targeting underrepresented groups, but they can also apply to the entire sample, such as stopping the data collection. This later application is the subject of this chapter.

The concept of adaptive designs originated in clinical trial studies with dynamic

treatment regimes, where different treatments are considered depending on characteristics of the patient and responses to previous treatments (Murphy, 2003). The goal is to increase response rates and avoid nonresponse bias by creating some decision rules to be applied to each individual in the sample. Wagner (2008) suggested applying some of these methods from dynamic regimes to survey methodology. Considering different strategies suited to each individual's features can increase their probability of response. This not only increases response rates, but also can help improve the representativeness of the sample.

In order to measure this representativeness, Schouten et al. (2009) proposed the concept of representative indicators, or R-indicators. These are an alternative to using nonresponse rates as quality measures of nonresponse bias. The R-indicators are based on estimates of the response probabilities and their distance to the value of a representative sample. Here, a sample is defined as representative if the response propensity is constant; that is, the respondents are a random subsample of the sample. To identify groups to improve the quality of the sample, Schouten et al. (2011) extend this concept to partial R-indicators. Here, representativeness is defined within subpopulations formed by auxiliary variables and the response propensity for each subgroup. The R-indicators and partial R-indicators can be used to monitor and tailor data collection, as well as compare waves in a longitudinal survey and different surveys on the same population (Schouten et al., 2012). Shlomo et al. (2012) discuss some methods for estimation of such indicators.

The idea of applying different design options at each survey phase is applied and exemplified by Groves and Heeringa (2006) in what they define as responsive survey design. These design options include: various means for applying the questionaries, like mail, telephone, internet and face-to-face interview; level of incentive offered to respond; number of follow-up calls; and the choice of short or long questionaries. Based on cost and error measures, and paradata collected at previous phases, the

agency can decide on what design options to apply in the next phases. This responsive design depends on the data collected in the first phase to guide the strategy forward. Responsive designs and other methods of applying adaptive design schemes are reviewed by Schouten et al. (2013). The main difference of the responsive designs (Groves and Heeringa, 2006) is that the measures are identified only during data collection in the first phases, while adaptive designs normally assume there is some prior information available to specify the strategies before starting the survey.

Data collected at initial phases also can be used to guide decisions applied to general features of the survey design, not only to individually tailored strategies. For example, the agency can decide to stop data collection depending on the amount of information already recorded and the costs of completing data collection. The benefits of stopping data collection include reducing the total cost and releasing results earlier, but it only makes sense if the data quality and inference are not sacrificed to undesirable levels. The quality of data can be assessed by information measures based on quantities of interest and their estimated changes at each survey phase. Rao et al. (2008) propose some stopping rules for surveys with multiple waves of follow-up for binary response variables. Their most robust measure is based on standardized differences of the response proportions at each wave, where the proportions are estimated with multiple imputation of the nonresponses. The nonrespondents' data are imputed from the predictive logistic regression model fitted to the observed data from previous waves. Wagner and Raghunathan (2010) also propose a stopping rule based on the probability of additional data changing the estimate, and compare their results with the rules proposed by Rao et al. (2008).

The criterion used in the stopping rule of Wagner and Raghunathan (2010) is based on the difference of two estimated proportions of the binary response variable. The first measure assumes that data collection stops and the nonrespondents are imputed from the available data, whereas the second measure assumes that an

extra wave of data is collected and a similar imputation approach is carried. In both cases, the nonrespondents are imputed under the assumption that they follow the same distribution as the respondents; that is, the data are missing at random (MAR). However, in some applications the nonresponse may be missing not at random (MNAR). Some imputation methods used to deal with this problem are reviewed in Chapter 1 and Chapter 2.

When analyzing data that are collected in waves, one alternative is to consider that the data distribution can change depending on the wave. Rao et al. (2008) try to take this into account with their proposed stopping rules by introducing a new scenario where the nonrepondents are imputed based only on the respondents of the last wave, instead of all previous waves. However, the data are still assumed to be MAR from the last wave. We proposed an imputation method in Chapter 2 where the user can impute data under MNAR by changing the mixture probabilities of the model fitted to the observed data. In this chapter, we describe how to use the method for MNAR imputation under an adaptive design perspective.

Consider an ongoing survey that is being evaluated at a certain point in time to decide whether it is worth continue collecting data, or to stop it and impute the nonrespondents from what has been observed so far. This decision depends on the quantity and quality of the data that were already recorded, how different the nonrespondents are from the respondents, and how any differences could impact the inference results. To facilitate such decision-making, we propose some utility measures to estimate the change in the results under different missingness scenarios. Each scenario reflects one possibility for the missingness pattern, ranging from the MAR case, where the imputed data are generated from the same distribution as the respondents, to other cases of nonignorable missingness. For each scenario considered by the analyst, we propose calculating the utility measures and the cost for various follow-up sample sizes. As the sample size increases, the utility measures

are naturally getting better, but the cost also increases. Thus, the decision rule will depend on the trade-off between these measures. With the values of utility and cost measures for different sample sizes and under different missingness scenarios, the agency can make an informed decision about stopping the data collection.

We describe our approach of adaptive design with imputation methods in Section 3.2. We briefly review the model and the imputation method in Section 3.2.1 and Section 3.2.2. We describe the steps to create the completed data sets to evaluate the adaptive design in Section 3.2.3. We define the utility and cost measures in Section 3.2.4 and Section 3.2.5, respectively. The method is exemplified by an application to the U.S. Census of Manufactures in Section 3.3. Additional discussion appears in Section 3.4.

3.2 Methodology

Applying an adaptive design to a large scale survey can improve the quality of the data while optimizing budget allocation. Depending on the representativeness of the respondents, it may be cost effective to stop data collection and impute the nonrespondents. At a certain point in time, this decision will depend on the impact on inference of the imputation values as they get further from being missing at random. Before discussing the steps necessary to make decisions about the extent of follow-up, we review the imputation method from Chapter 2 that will be used in different stages of the survey.

3.2.1 Mixture model

Imputation of the missing data is an important part of the adaptive design approach we propose, as we generate completed survey data under different scenarios. These scenarios can include Missing at Random (MAR) data, as well as Missing Not at Random (MNAR). The model has also to be flexible to capture different patterns of

the multivariate continuous data, such as multi modality, skewness and correlation. The mixture of multivariate normal distributions is a natural choice, since with a sufficient number of components, it can approximate well any distribution. With a nonparametric prior in a Bayesian framework (Ferguson, 1973, 1983; Escobar and West, 1995; West et al., 1994), we can provide more flexibility and improve the density estimation (Müller and Mitra, 2013).

Let $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)'$ denote the p variables of the n respondents. Denote by $z_i \in \{1, \ldots, K\}$ the indicator of which component the i-th observation belongs to with probability $\pi_k = P(z_i = k)$, where $i = 1, \ldots, n$ and $k = 1, \ldots, K < \infty$. Within each component, $\boldsymbol{Y}$ follows a multivariate normal distribution with mean $\boldsymbol{\mu}_k$ and variance Σ_k:

$$\begin{aligned} \boldsymbol{y}_i | z_i, \boldsymbol{\mu}, \boldsymbol{\Sigma} &\sim N(\boldsymbol{\mu}_{z_i}, \Sigma_{z_i}) && (3.1) \\ z_i | \boldsymbol{\pi} &\sim \text{Multinomial}(\pi_1, \ldots, \pi_K). && (3.2) \end{aligned}$$

We standardize each dimension of $\boldsymbol{Y}$ to facilitate modeling. In Chapter 2, we discuss two alternatives for the prior specification of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The first is the conjugate prior, with

$$\begin{aligned} \boldsymbol{\mu}_k | \Sigma_k &\sim N(\boldsymbol{\mu}_0, h^{-1}\Sigma_k) && (3.3) \\ \Sigma_k &\sim \text{InverseWishart}(f, \Phi), && (3.4) \end{aligned}$$

where f is the degrees of freedom and $\Phi = diag(\phi_1, \ldots, \phi_p)$ with $\phi_j \sim \text{Gamma}(a_\phi, b_\phi)$ for $j = 1, \ldots, p$. We set $\boldsymbol{\mu}_0 = 0$, since the variables are standardized, $f = p + 1$ to ensure a proper posterior distribution, and $h = 1$ for convenience. The second alternative has the same prior as (3.3) for $\boldsymbol{\mu}$, but instead we set the covariance matrices to $\Sigma_k = \sigma I_p$, for all k and for a specified value of $\sigma > 0$. The value of σ controls the tightness of the clusters. In the remaining of this chapter, we use this approach with the fixed covariances, since the results in Chapter 2 suggest this

approach provides more flexibility for the imputation models.

Following the stick-breaking representation of a truncated Dirichlet process (Sethuraman, 1994; Ishwaran and James, 2001), we define the mixture weights as

$$\pi_k = v_k \prod_{g<k}(1 - v_g) \quad \text{for } k = 1, \ldots, K \tag{3.5}$$

$$v_k \sim \text{Beta}(1, \alpha) \quad \text{for } k = 1, \ldots, K-1;\ v_K = 1 \tag{3.6}$$

$$\alpha \sim \text{Gamma}(a_\alpha, b_\alpha). \tag{3.7}$$

The hyperparameters are set following the same specification of Kim et al. (2014), with $a_\alpha = b_\alpha = 0.25$ to allocate the probabilities to the first few components. More details about the model and another alternatives for the prior specification are discussed in Chapter 2.

3.2.2 Imputation methods

The model in (3.1)–(3.7) can be used to impute values for the missing observations. The imputed values are simulated from the posterior predictive distribution within the Gibbs sampler. If we use the estimated values of $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ on the predictive distribution, the imputed values are generated from the same distribution as the observed data. This is done when the missing data are assumed to be missing at random. As we proposed in Chapter 2, we can use the estimated $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and specify new mixture probabilities $\boldsymbol{\pi}^*$ to generate imputed values from a distribution different than the observed. This is done when the missing data are assumed to be missing not at random, and $\boldsymbol{\pi}^*$ reflects the assumptions about the missingness pattern.

The model also can be used to fill in the missing data caused by item nonresponse, when some of the variables are observed, and by unit nonresponse, when none of the variables are observed. In both cases, $\boldsymbol{y}$ is sampled from the conditional distribution given the posterior samples of the parameters. The parameters are sampled from their corresponding full conditionals, as described in Section 2.2.1 of Chapter 2.

Item nonresponse

For item nonresponse, we impute the missing variables for each unit from the conditional normal distribution given the observed variables. For this case, we assume that the cause for nonresponse is not related to the missing responses, that is, the item nonresponses are missing at random given the observed variables.

Denote by $(\boldsymbol{\pi}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ the posterior samples of the parameters with $t = 1, \ldots, T$ for T iterations. At each MCMC iteration, we first update $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from their full conditionals. Then, for $i = 1, \ldots, n$, we update

$$z_i^{(t)} | \boldsymbol{y}_i^{(t-1)}, \boldsymbol{\pi}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)} \sim \text{Multinomial}(\pi_{i1}^*, \ldots, \pi_{iK}^*), \tag{3.8}$$

where $\pi_{ik}^* = \pi_k^{(t)} N(\boldsymbol{y}_i^{(t-1)} | \boldsymbol{\mu}_k^{(t)}, \Sigma_k^{(t)}) / \left\{ \sum_{g=1}^K \pi_g^{(t)} N(\boldsymbol{y}_i^{(t-1)} | \boldsymbol{\mu}_g^{(t)}, \Sigma_g^{(t)}) \right\}$, and $\boldsymbol{y}_i^{(t-1)}$ is the completed vector with the observed and imputed parts from the previous iteration for the i-th observation.

Let $\boldsymbol{y}_i = (\boldsymbol{y}_{i,\text{obs}}, \boldsymbol{y}_{i,\text{mis}})$ denote the partition of the p-dimensional vector of the i-th observation into the observed and missing parts. Given the component indicator z_i, the mean and covariance matrix are also partitioned and reordered into the corresponding observed and missing parts as

$$\boldsymbol{\mu}_{z_i}^{(t)} = (\boldsymbol{\mu}_O, \boldsymbol{\mu}_M) \qquad \text{and} \qquad \Sigma_{z_i}^{(t)} = \begin{pmatrix} \Sigma_{OO} & \Sigma_{OM} \\ \Sigma_{MO} & \Sigma_{MM} \end{pmatrix}. \tag{3.9}$$

Then, we update $\boldsymbol{y}_{i,\text{mis}}^{(t)}$ from

$$\boldsymbol{y}_{i,\text{mis}}^{(t)} | \boldsymbol{y}_{i,\text{obs}}, z_i^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)} \sim N\left(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}\right), \tag{3.10}$$

where

$$\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}_M + \Sigma_{MO} \Sigma_{OO}^{-1} (\boldsymbol{y}_{i,\text{obs}} - \boldsymbol{\mu}_O) \tag{3.11}$$

$$\bar{\Sigma} = \Sigma_{MM} - \Sigma_{MO} \Sigma_{OO}^{-1} \Sigma_{OM}. \tag{3.12}$$

Unit nonresponse

For unit nonresponse, we impute the entire vector of variables for the n_{mis} completely missing respondents from the posterior predictive distribution. In this case, there is no partial information about the location of the responses. Here, we assume that the values are missing not at random. As in Chapter 2, we use the samples of $\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\Sigma}^{(t)}$ from the model fit to the observed data, and specify new mixture probabilities $\boldsymbol{\pi}^*$. For $i = 1, \ldots, n_{\text{mis}}$, we sample

$$z_i|\boldsymbol{\pi}^* \sim \text{Multinomial}(\pi_1^*, \ldots, \pi_K^*) \tag{3.13}$$

$$\tilde{\boldsymbol{y}}_i|z_i, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)} \sim N(\boldsymbol{\mu}_{z_i}^{(t)}, \Sigma_{z_i}^{(t)}). \tag{3.14}$$

Since we need to identify the components to choose the new probabilities, we propose ranking the clusters post-simulation based on their distance δ to the origin, i.e., $\delta_{\text{orig}} = \boldsymbol{\mu}'\boldsymbol{\mu}$, or to the minimum value with $\delta_{\text{min}} = (\boldsymbol{\mu} - \boldsymbol{y}_{\text{min}})'(\boldsymbol{\mu} - \boldsymbol{y}_{\text{min}})$, where $\boldsymbol{y}_{\text{min}} = (\min_i(\boldsymbol{y}_1), \ldots, \min_i(\boldsymbol{y}_p))$, with $\min_i(\boldsymbol{y}_v) = \min(y_{1v}, \ldots, y_{nv})$ for all variables $v = 1, \ldots, p$. Thus, if we want to impute more data with higher response values, we can inflate the probabilities of the top ranked clusters, and the reverse if we want to generate more data with lower response values.

We also summarize the posterior samples by choosing the MCMC iteration with largest posterior value (Fraley and Raftery, 2007). For each iteration, we evaluate the posterior given the sampled values for the parameters and the sample of the complete observed and imputed $\boldsymbol{Y}$. Then, we select the iteration with maximum posterior value (MAP) to summarize the samples. This is done to simplify the task of setting $\boldsymbol{\pi}^*$ to just one cluster allocation, since this allocation can change from iteration to iteration.

After choosing the MAP iteration and ranking the components, we recommend using the NIMC (Nonignorable missingness Imputation for Multivariate Continuous

data) application to choose the probabilities $\boldsymbol{\pi}^*$. We describe and demonstrate the implementation of the NIMC application in Chapter 2.

3.2.3 Adaptive Design

Let us consider an ongoing survey which is going to be evaluated at a certain point in time. The goal is to decide if it is worth spending any more resources collecting more data, or if what has been collected so far is adequate to enable inferences for population quantities. This decision is based on the trade-off between the improvement in the data representativeness and the increase in cost of collecting more data.

We denote by N the intended size of the entire sample. After the first wave of data collection, this sample is partitioned into the respondents, who have at least one of the variables recorded, and the nonrespondents, who did not provide any response to the survey yet. To deal with the item nonresponse among the respondents, we follow the first approach described in Section 3.2.2 and impute the missing values assuming MAR conditioned on the observed values. We assume that this step is done within the MCMC whenever we refer to fitting the mixture model to the respondents. This step does not change depending on the decision about the adaptive design.

We consider two options to obtain a completed data set of size N. The first option is to stop collecting data and impute the missing data based on the information available from the observed data. We fit the multivariate mixture model described in Section 3.2.1 to the n_R respondents denoted by D_R. The remaining n_{NR} nonrespondents in set D_{NR} are then imputed based on this model, as seen in Figure 3.1. The imputation at this stage can be done under scenarios of either MAR or MNAR, depending on the purpose of the completed data. When generating a completed data set to be released for general analysis, the missing data are probably going to be imputed assuming MAR. When doing sensitivity analysis to evaluate the impact of missingness patterns, we consider different possible scenarios, including MNAR.

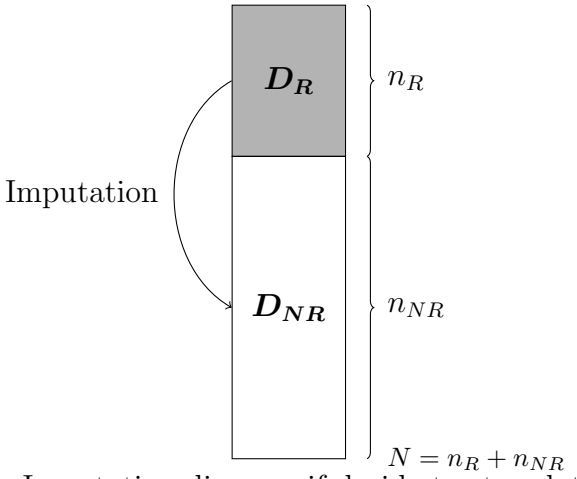


FIGURE 3.1: Imputation diagram if decide to stop data collection

We show some different scenarios in the results in Section 3.3.

The second option is to collect n_F more observations to form a follow-up sample, D_F. The set of n_{NF} cases that are still missing after the follow-up sample are denoted by D_{NF}. The mixture model is fit again, either to all the observations in $(D_R + D_F)$ or just D_F, as seen in Figure 3.2(a) and 3.2(b) respectively. These alternatives correspond to the belief that the nonrespondents in D_{NF} are more similar to the entire set of observations or just to the latest wave. It may be beneficial to use $(D_R + D_F)$ when n_F is too small to support reliable modeling.

The choice in Figure 3.2(a) is appropriate if D_R and D_{NR} come from the same distribution and the latter is MAR. Here, D_F is a random sample of the nonrespondents, and D_{NF} should be imputed from the same distribution as all the data that was observed so far. It also makes sense to follow 3.2(a) if D_R and D_{NR} have different distributions, but D_F comes from the same distribution of D_R, and the differences in the distributions are reflected on a MNAR imputation model. The choice in Figure 3.2(b) makes more sense if it is believed that D_R and D_{NR} come from different distributions, D_F is a random sample of the nonrespondents, and D_{NF} should be imputed based only on the distribution of the follow-up sample.

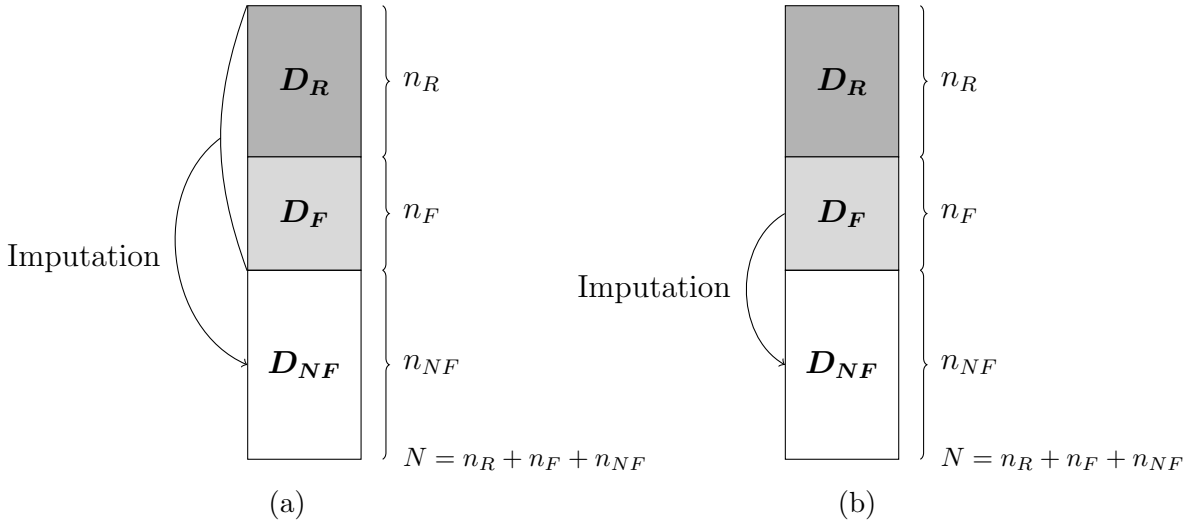


FIGURE 3.2: Imputation diagrams if decide to collect follow-up sample

The decision between the two options represented in Figures 3.1 and 3.2 depends on the trade-off between utility and cost of collecting more data. In order to evaluate the utility of what has been collected so far, we propose comparing some utility measures for different follow-up sample sizes and under different missingness scenarios via sensitivity analysis.

Before defining the utility measures, we describe the steps to create completed data sets under the different imputation options. With the data collected on the first survey wave, we fit the multivariate mixture model from Section 3.2.1 to the observed values. After the convergence of the MCMC and the selection of the iteration with maximum posterior value, we summarize the model estimate with the selected MAP posterior samples of the mixture parameters. The imputed values due to item nonresponse are also selected from the MAP iteration. Combined with the actual recorded values, they form the set D_R that is going to be used in all the imputation scenarios.

With the fitted cluster allocation, the next step is to consider some plausible missingness scenarios for sensitivity analysis. This is done by specifying new prob-

abilities $\boldsymbol{\pi}^*$ to reflect different distributional patterns for the nonrespondents. To facilitate this step, we developed the NIMC (Nonignorable missingness Imputation for Multivariate Continuous data) application. The NIMC application provides an interface where the user can set the values of $\boldsymbol{\pi}^*$ with sliders for each component, and immediately see pairwise scatterplots and summary statistics of the observed and imputed data.

The scenarios can include the MAR case as a baseline, as well as other MNAR cases that are reasonable for the survey context. An example of a MNAR case is when the probabilities in $\boldsymbol{\pi}^*$ are increased for the top ranked clusters to reflect a larger proportion of missing data with larger responses.

These scenarios are used to compare the impact on inferences of different sizes of follow-up samples, starting from what has been observed so far, and progressing as the follow-up sample size increases. As n_F increases, inferences on the MAR case are not expected to change as much as on the MNAR case. If the impact on the inferences is significant, the agency might decide that it is worth collecting more data. If even in an extreme MNAR scenario the impact is not significant, then the agency might decide to stop data collection.

We describe now the data sets that will be used to evaluate these impacts. Define n_{MAX} as the maximum sample size that can be collected given a total budget or a time restriction. We consider sampling different proportions of the maximum sample size, such that $n_F = \delta \times n_{MAX}$ where n_F is the follow-up sample size and $\delta \in [0, 1]$. Define m_P as the number of multiple imputations we generate when imputing to create the "population" according to the missingness scenario, and m_F as the number of multiple imputations we generate when imputing the remaining data after the follow-up sample.

We denote by $\Omega = (\boldsymbol{\pi}^{*(1)}, \ldots, \boldsymbol{\pi}^{*(S)})$ the set of probabilities of each scenario to be considered. For each value of $\boldsymbol{\pi}^{*(s)} \in \Omega$, with $s = 1, \ldots, S$, we do the following

procedure:

1. Generate m_P completed hypothetical populations $(P^{(s,1)}, \ldots, P^{(s,m_P)})$ by multiply imputing all the non-respondents with the fitted model and the probabilities $\boldsymbol{\pi}^{*(s)}$. The observed values D_R are common to all s. The imputed values for each hypothetical population are denoted by $\tilde{D}_{NR}^{(s,j)}$, such that $P^{(s,j)} = D_R \cup \tilde{D}_{NR}^{(s,j)}$, for $j = 1, \ldots, m_P$.

2. For each value of δ under consideration, and for each $j = 1, \ldots, m_P$:

   (a) Obtain a random sample of size n_F from each $\tilde{D}_{NR}^{(s,j)}$. Denote this follow-up sample set by $D_{F,\delta}^{(s,j)}$.

   (b) Fit new mixture models to one or both:

      i. $D_R \cup D_{F,\delta}^{(s,j)}$

      ii. $D_{F,\delta}^{(s,j)}$

      depending on the option selected from Figure 3.2(a) and Figure 3.2(b).

   (c) Based on the posterior samples of the new mixture models and assuming MAR, generate m_F multiply imputed values for the n_{NF} observations that are still missing. The imputed cases are denoted by $\tilde{D}_{NF,\delta}^{(s,j,l)}$, with $l = 1, \ldots, m_F$. The completed data sets are denoted by $\tilde{D}_{\delta}^{(s,j,l)} = D_R \cup D_{F,\delta}^{(s,j)} \cup \tilde{D}_{NF,\delta}^{(s,j,l)}$.

   (d) Compare the sets $P^{(s,j)}$ and $\left(\tilde{D}_{\delta}^{(s,j,1)}, \ldots, \tilde{D}_{\delta}^{(s,j,m_F)}\right)$ with the utility measures described next.

We generate these multiple draws to account for the uncertainty due to the missing data on each stage of the process. The uncertainty from selecting the follow-up sample could also be considered with repeated sampling of $D_{F,\delta}^{(s,j)}$. However, this

source of uncertainty is already accounted for with a reasonable number m_P of multiple imputations.

3.2.4 Utility measures

The decision of stopping the data collection depends critically on the amount of information that has been collected so far. This information can be measured by the difference in inferences based on what was observed and inferences with a larger sample, under different missingness hypothesis. If the observed sample size is small and the non-respondents have an extremely different distribution than what was recorded, then the impact of a follow-up sample on the estimates is going to be bigger than if the initial sample was larger or the non-respondents were closer to be missing at random.

Consider the two complete data sets generated in Section 3.2.3: $P^{(s,j)}$, the hypothetical population for scenario s and imputation j; and each $\tilde{D}_\delta^{(s,j,l)}$, the completed data set considering the follow-up sample for imputation l. We consider the following measures to compare the similarities between $P^{(s,j)}$ and $\tilde{D}_\delta^{(s,j,l)}$.

Measure τ:

Let $\bar{Y}_v^{P(s,j)} = \sum_{i=1}^N Y_{i,v}^{P(s,j)}/N$ denote the marginal mean of each variable $v = 1, \ldots, p$ computed from the data set $P^{(s,j)}$, and let $\bar{Y}_v^{D,\delta(s,j,l)} = \sum_{i=1}^N Y_{i,v}^{D,\delta(s,j,l)}/N$ denote the marginal mean of each variable $v = 1, \ldots, p$ computed from the data set $\tilde{D}_\delta^{(s,j,l)}$. Similarly, let $\hat{\sigma}_v^{P(s,j)}$ and $\hat{\sigma}_v^{D,\delta(s,j,l)}$ denote the standard deviations for each variable v from the data sets $P^{(s,j)}$ and $\tilde{D}_\delta^{(s,j,l)}$, respectively. For each variable $v = 1, \ldots, p$, compute

$$t_v^{\delta(s,j,l)} = \frac{\left(\bar{Y}_v^{P(s,j)} - \bar{Y}_v^{D,\delta(s,j,l)}\right)}{\sqrt{\frac{\left[\left(\hat{\sigma}_v^{P(s,j)}\right)^2 + \left(\hat{\sigma}_v^{D,\delta(s,j,l)}\right)^2\right]/2}{N}}}. \tag{3.15}$$

Let the summary measure be

$$\tau^{\delta(s,j,l)} = \frac{1}{p}\sum_{v=1}^{p}\left|t_v^{\delta(s,j,l)}\right|, \tag{3.16}$$

for each value of δ, scenario s, population j, and imputation l.

Measure θ:

We also define a measure based on the Mean Absolute Percentage Error, a measure used to measure the percentage error between forecasts and observed values. Here, we consider the average percentage difference between $\bar{Y}_v^{P(s,j)}$ and $\bar{Y}_v^{D,\delta(s,j,l)}$, with respect to $\bar{Y}_v^{P(s,j)}$. Thus, the summary measure is

$$\theta^{\delta(s,j,l)} = \frac{1}{p}\sum_{v=1}^{p}\left|\frac{\bar{Y}_v^{P(s,j)} - \bar{Y}_v^{D,\delta(s,j,l)}}{\bar{Y}_v^{P(s,j)}}\right|, \tag{3.17}$$

for each value of δ, scenario s, population j, and imputation l.

Measure ρ:

Since these measures are focused on the marginal differences, characteristics of the multivariate data may not be captured. Woo et al. (2009) propose global measures of data utility to compare multivariate distributions of two data sets, in an attempt to quantify the dissimilarity between a masked data set and an original confidential data set that cannot be released. They found that a measure based on propensity scores is the most promising to reflect characteristics of the entire distribution for different types of data, while being computationally feasible to implement.

Propensity scores are used in observational studies for matching covariate characteristics and reduce the impact of confounding factors between groups when inferring treatment effects. The propensity score is defined as $e(\boldsymbol{x}) = P(T = 1|\boldsymbol{x})$, the probability of being assigned to be on the treatment group T given the variables $\boldsymbol{x}$. Woo

et al. (2009) propose calculating the propensity score on the merged data set consisting of the original and the masked data. The treatment response T is the indicator for the synthetic data, and the propensity score is estimated for each unit in the merged data set. The similarity between the two groups is assessed by the distribution of the estimated propensity scores and how close they are to 0.5, the reference for indistinguishable groups.

The model specified to estimate the propensity scores has a great impact on this measure. The common approach is to use simple logistic regression of the treatment indicator on the covariates, which assumes a linear relationship between the link function and the covariates. Woo et al. (2008) propose a more flexible approach by using generalized additive models (GAM) instead of the standard logistic regression. The linear component of the regression is replaced by a flexible additive function:

$$\text{logit}\Big(e(\boldsymbol{x})\Big) = \log\left(\frac{e(\boldsymbol{x})}{1-e(\boldsymbol{x})}\right) = \alpha + f_1(x_1) + \cdots + f_p(x_p), \tag{3.18}$$

where each $f_j(x_j)$ is a smooth function of x_j, for example regression splines. This model outperforms the logistic regression and facilitates the modeling of non-liner relationships. However, as in the logistic regression, the results are dependent on the variables and interactions that are included in the model. The recommendation remains to include the covariates believed to be related to the "treatment" indicator.

We define the measure ρ based on the propensity score (Woo et al., 2009) estimated with GAM (Woo et al., 2008), with the covariates as all the response variables $\boldsymbol{y} = (y_1, \ldots, y_p)$ included as main effects. The two data sets to be compared, $P^{(s,j)}$ and $\tilde{D}_\delta^{(s,j,l)}$, are merged together. The set $P^{(s,j)}$ receives an indicator variable $T = 1$, while $\tilde{D}_\delta^{(s,j,l)}$ has $T = 0$. The logistic model is fit on the response variable T, for the merged data set of size $2N$. The predicted values of each observation $\hat{e}_i$ should be

around 0.5 if the two data sets are comparable. The overall measure is given by

$$\rho^{\delta(s,j,l)} = \frac{\sum_{i=1}^{2N}(\hat{e}_i - 0.5)^2}{2N}, \tag{3.19}$$

for each value of δ, scenario s, population j, and imputation l. If the distributions are equal, ρ is close to zero. On the other extreme, if the two distributions are very distinct, then $\rho \sim 1/4$.

As n_F increases, the two data sets intersect more and, obviously, all the measures will decrease. Therefore, the interest is on their relative values compared between the different follow-up sample sizes and for the different scenarios. Together with the cost measures, the agency conducting the survey can make a decision about stopping data collection.

3.2.5 Cost measure and decision rule

Together with the utility measures described above, with a measure cost, one can make the decision about the follow-up sample and its size. We propose obtaining the results with a set of different values of n_F, calculating the utility measures (τ, θ, ρ) and estimating the cost of collecting the additional sample. With this information, the agency can decide how much it is willing to spend for the utility improvements. The values can be explored with a table or plot, as we show in Section 3.3.

Denote by C_F the total cost of the follow-up sample, C_0 the fixed cost regardless of the sample size, and c the cost of selecting, measuring and processing each of the n_F follow-up sample units. For our purposes of creating a decision rule for adaptive design, we consider sufficient to estimate the cost with a linear function of the sample size. More complex cost functions are discussed in Groves (2004). The total cost is estimated by

$$C_F = C_0 + (c \times n_F). \tag{3.20}$$

We note that the maximum sample size n_{MAX}, defined and used in Section 3.2.3, can be obtained by solving for n_F in (3.20) given the total budget available.

3.3 Illustration with Census of Manufactures Data[1]

We demonstrate now the proposed approach using data from the 2007 U.S. Census of Manufactures (CMF). The original data consist of responses collected from forms sent to companies representing all U.S. locations and industries. The form queries include information about sales, employees and payroll of the businesses. As in many other large scale surveys, the CMF faces the problem of missing data, from both item and unit nonresponse. Because of this, the U.S. Census Bureau spends resources trying to reduce the nonresponse rates during data collection, for example, by resending the forms, or allocating Census consultants to work specifically with some of the largest companies. Thus, the CMF could potentially benefit from an adaptive survey design to stop data collection earlier. This would not only reduce the cost of the survey, but it can also be beneficial for users with data released sooner.

The CMF was the main motivating case for the proposed adaptive design methodology with nonignorable missingness. This is because the data are collected and processed in waves during the year, depending on when each firm sends its form. With this temporal pattern, it is natural to think about evaluating the survey design at certain time points, e.g., monthly or quarterly. This pattern is also a mechanism that can lead to missing not at random data. It can happen, for example, if smaller companies tend to not send their forms on time. If that is the case, with our imputation model, we can inflate the probabilities of bottom-ranked clusters to impute more data with lower response values. This and some other scenarios are reasonable assumptions to consider when making sensitivity analysis, as we demonstrate here.

[1] DISCLAIMER: Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed.

In order to investigate the missingness patterns and apply our imputation method, we focus on some selected industries and variables of the CMF. The variables are total value of shipments (TVS), total employment (TE), and salary/wages (SW). The first selected industry is ready-mix concrete manufacturing, which is a homogeneous industry since most of their business is based on a single product. The second selected industry is plastics products manufacturing, which is a more broad classification and more heterogeneous.

For illustration, we consider as observed data all the units that had at least one reported value among the three variables of interest and that had a valid response date. Some observations met the first criterion, but there was no response date available on file for miscellaneous reasons, which could include the case of late respondents that sent their information after some processing deadline. We consider these cases as missing data, and impute their entire vector of responses after fitting the mixture model in (3.1)–(3.7) to the observed data.

For each industry, we selected three scenarios to consider for the sensitivity analysis. The first is the MAR case, with the estimated probabilities $\boldsymbol{\pi}$, as a baseline. The second is the MNAR case with higher probabilities for bottom ranked clusters, to generate small response values. The third scenario is the opposite MNAR case, with higher probabilities for top ranked clusters. This is included in the sensitivity analysis as the worst case scenario, since larger companies will have more impact in some summary statistics. With the NIMC application, we specified the values of the probabilities and obtained $\Omega = (\boldsymbol{\pi}^{*(\text{MAR})}, \boldsymbol{\pi}^{*(\text{bottom})}, \boldsymbol{\pi}^{*(\text{top})})$ corresponding to three scenarios described above. The values of $\boldsymbol{\pi}^*$ for each scenario can be seen in the legends of Figure 3.3, Figure 3.4 and Figure 3.5 for the concrete industry, and in Figure 3.7, Figure 3.8 and Figure 3.9 for the plastic industry.

We considered $\delta = (0, 0.25, 0.5, 0.75, 1)$, that is, collecting follow-up samples with 0%, 25%, 50%, 75% and 100% of n_{MAX}. For this example, we set $n_{MAX} = n_{NR}$,

assuming the budget and time available are enough to attempt to collect all non-respondents. For each industry and for each scenario, first we created $m_P = 10$ complete hypothetical populations. Then, for each population $P^{(s,j)}$ and for each value of δ, we obtained the follow-up sample $D_{F,\delta}^{(s,j)}$. We fit two new mixture models to: (i) the observed data set and the follow-up sample $\left(D_R \cup D_{F,\delta}^{(s,j)}\right)$; and to (ii) the follow-up sample only $\left(D_{F,\delta}^{(s,j)}\right)$. With the posterior samples from the new mixture models, we create $m_F = 5$ complete data sets through multiple imputation of the still remaining nonrespondents. For all scenarios, we calculated the utility measures ρ, τ and θ as described in Section 3.2.4. Each measure is summarized by the mean and standard deviation of the individual values from m_P population repetitions and the m_F multiple imputations.

We do not include values for the cost measure due to disclosure limitations of the CMF data, per request of the Census Bureau. The cost measure in (3.20) is a linear function of n_F, and thus, a linear function of δ. Thus, we believe it is sufficient to present the results based on δ for the purpose of sensitivity analysis and comparison of the different scenarios.

In Figure 3.3, we see the pairwise scatterplots of the results of the MAR scenario for the concrete industry. The axes of all the scatterplots were removed to prevent disclosure of information about the magnitude of the data. On the concrete data, the mixture model resulted in 17 nonempty components plotted by the colored circles in the upper diagonal plots. From the estimated probabilities in the legend, we can see that most of the weight is concentrated in less than half of the clusters. The black points in the lower diagonal plots are imputed assuming MAR. In Figure 3.4, we can see the imputation results of the second scenario, with higher probabilities for bottom ranked clusters for the concrete industry. Because of this pattern on the probabilities, there are more imputed points on the lower tails of the data. In

Figure 3.5, we can see the imputation results of the opposite scenario, with higher probabilities for the top ranked clusters for the concrete industry.

In Table 3.1, we can see the utility measures of the MAR scenario for the concrete industry, with the two approaches for the new mixture model. In Table 3.2, we have the results for the MNAR scenario with higher probabilities for bottom ranked clusters, and in Table 3.3, we have the results for the MNAR scenario with higher probabilities for top ranked clusters. As expected, in all cases the measures decrease as δ increases. When $\delta = 1$, we sample all the units and there is no need to fit the model again and impute more data. Thus, the two populations being compared, $P^{(s,j)}$ and $\tilde{D}_{\delta}^{(s,j,l)}$, are the same and the utility measures are equal to zero. When $\delta = 0$, we do not collect any more data, so there is no follow-up sample to fit the model on the second approach.

The MAR scenario is the baseline and the measures are relatively small, since the two data sets being compared should come from very similar distributions. In the MNAR scenarios, the values from the model fit with just the follow-up sample in Table 3.2(b) and Table 3.3(b) are smaller than the values from the model fit with the observed data. This is expected, since the distribution of the nonrespondents of the first wave, plotted as the black points in Figure 3.4 and Figure 3.5, is very different than the distribution of the respondents, plotted as the gray points. The second approach, therefore, creates data sets $\tilde{D}_{\delta}^{(s,j,l)}$ that are more similar to the populations $P^{(s,j)}$.

The measures are also compared with the plots in Figure 3.6. In these plots, we can compare the decrease of the measures between the different scenarios. The model fit with the observed data and the follow-up sample, plotted on the left column, resulted in very distinct measures for the scenarios. On the right column, the model fit with just the follow-up sample resulted in overlapping measures that are much smaller than with the first option. This confirms that the two options are generating

different completed data sets, and that using the model with just the follow-up sample is better when the data are MNAR. The values of θ, the measure based on the mean absolute percentage error, with the model fit with just the follow-up sample are larger for the MAR scenario than for the other scenarios. This could be caused by the instability of the percentage error when the values are close to zero, one of the drawbacks of the mean absolute percentage error. If the data are MAR, there is not much gain with a follow-up sample. If the data are MNAR, there is a significant utility improvement in collecting 25-50% more data. Beyond that, the utility increase might be not worth the increase in the cost.

With the utility measures and the cost of the follow-up samples, the agency should be able to make a decision about the survey. However, we can also make a decision based on a more formal method. As an example, let us consider the sample size that minimizes the sum of variance plus cost, as suggested by Groves (2004). In our case, we replace the variance by one of the utility measures, so that we minimize the measure of how far the imputed data are from the population. We choose the measure $\bar{\rho}$ to capture this overall distance between the data sets better, since the other measures are based on the distance between the marginal means. For the cost, we will use δ.

To make the variance and cost be on the same scale, Groves (2004) transforms the cost by multiplying it by a constant λ. Since in our illustrative example the proxy for cost, δ, is already on a fixed scale from 0 to 1, we will transform the values of $\bar{\rho}$ to be on the same scale by subtracting the minimum and dividing by the maximum at each scenario. Then, we select the follow-up sample size that minimizes the sum of this transformed $\bar{\rho}$ and δ. These cases are highlighted in bold in Table 3.1, Table 3.2 and Table 3.3. We note that under this transformation, the extreme cases of $\delta = 0$ and $\delta = 1$ are equivalent. The agency can consider other decision rules to guide the next stages of the survey.

Table 3.1: Summary of utility measures with the results for the Concrete industry from the 2007 CMF, for the MAR imputation scenario. The results highlighted in bold correspond to the follow-up sample size that minimized the sum of cost and the transformed measure $\bar{\rho}$. *The values of $\bar{\rho}$ (sd) should be multiplied by 10^{-5}.

(a) New mixture model fit to observed data and follow-up sample $\left(D_R \cup D_{F,\delta}^{(\text{MAR})}\right)$

δ	$\bar{\rho}$ $(sd)^*$	$\bar{\tau}$ (sd)	$\bar{\theta}$ (sd)
0	3.465 (1.823)	0.372 (0.242)	0.255 (0.229)
0.25	**2.514 (1.258)**	**0.290 (0.139)**	**0.185 (0.110)**
0.5	1.970 (1.045)	0.235 (0.119)	0.139 (0.082)
0.75	0.886 (0.348)	0.144 (0.091)	0.094 (0.075)
1	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)

(b) New mixture model fit to follow-up sample only $\left(D_{F,\delta}^{(\text{MAR})}\right)$

δ	$\bar{\rho}$ $(sd)^*$	$\bar{\tau}$ (sd)	$\bar{\theta}$ (sd)
0	–	–	–
0.25	6.858 (3.299)	0.424 (0.273)	0.279 (0.183)
0.5	3.662 (1.547)	0.313 (0.144)	0.224 (0.123)
0.75	**1.251 (0.573)**	**0.188 (0.092)**	**0.130 (0.074)**
1	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)

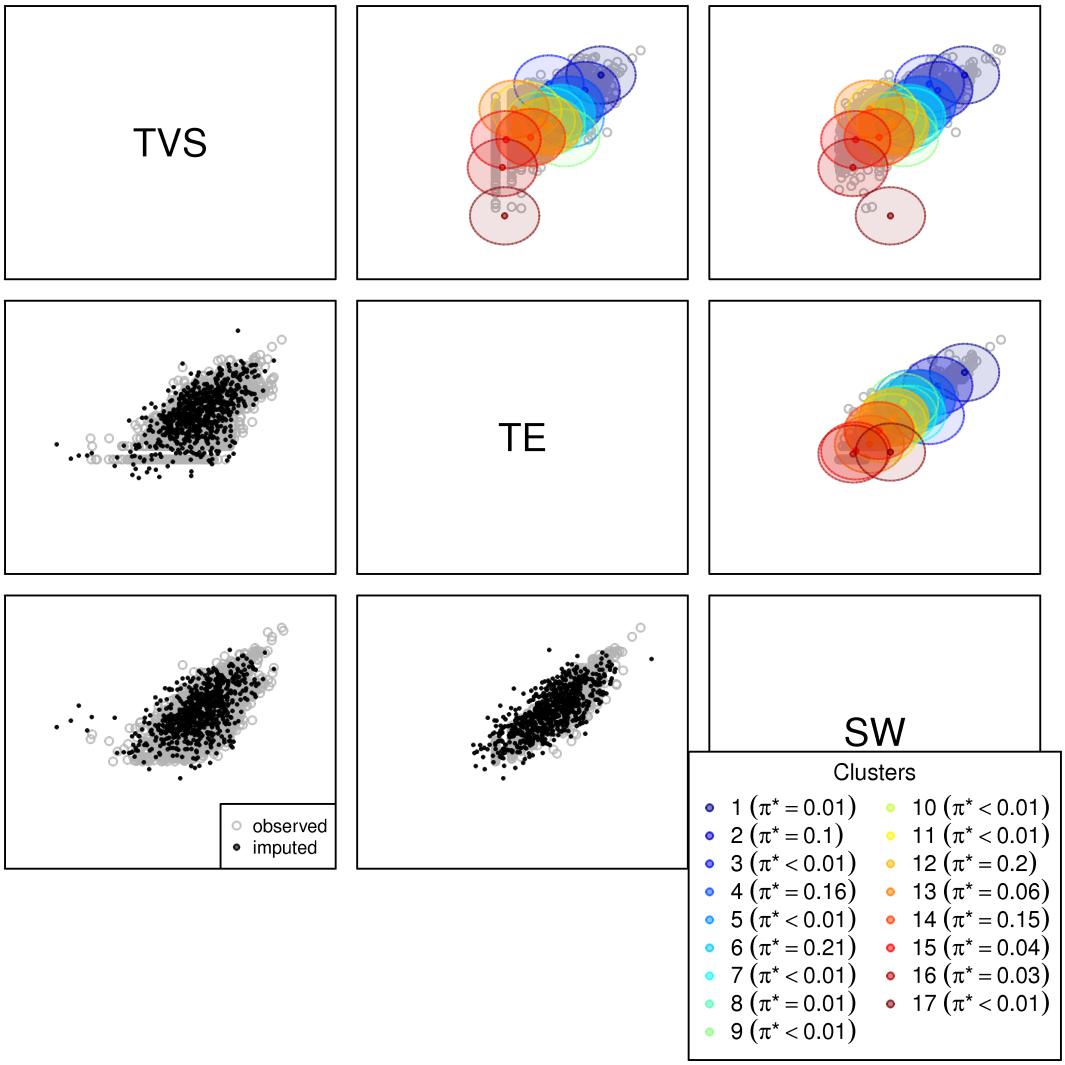


FIGURE 3.3: Pairwise scatterplots with the results for the Concrete industry from the 2007 CMF, for the MAR imputation scenario. Observed points are plotted as gray hollow circles. The black filled circles on the lower diagonal are the imputed points. The colored circles on the upper diagonal are the 95% quantile ellipses of the fitted clusters, with color intensity proportional to the mixture probabilities.

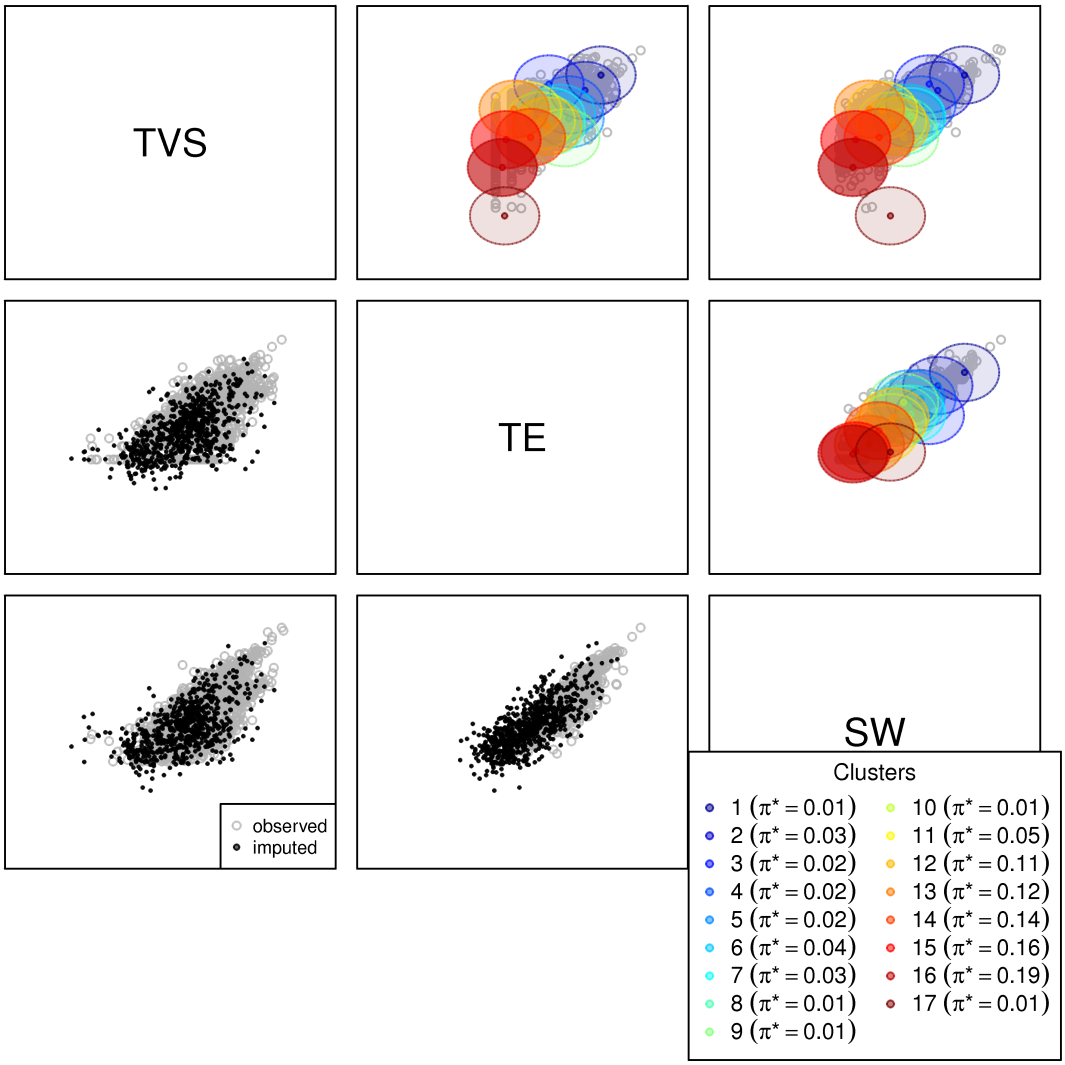


FIGURE 3.4: Pairwise scatterplots with the results for the Concrete industry from the 2007 CMF, for the MNAR imputation scenario with higher probabilities for bottom ranked clusters. Observed points are plotted as gray hollow circles. The black filled circles on the lower diagonal are the imputed points. The colored circles on the upper diagonal are the 95% quantile ellipses of the fitted clusters, with color intensity proportional to the mixture probabilities.

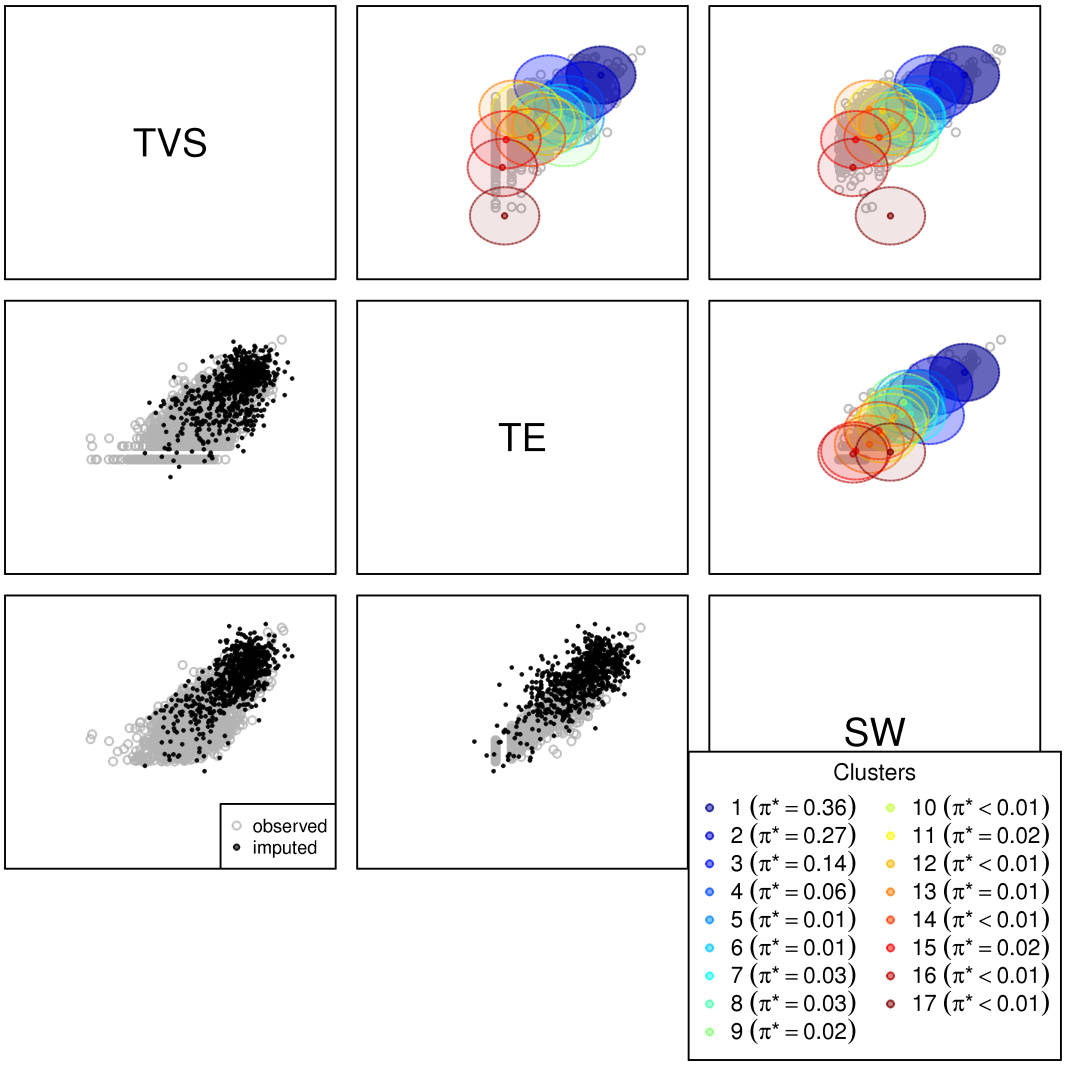


FIGURE 3.5: Pairwise scatterplots with the results for the Concrete industry from the 2007 CMF, for the MNAR imputation scenario with higher probabilities for top ranked clusters. Observed points are plotted as gray hollow circles. The black filled circles on the lower diagonal are the imputed points. The colored circles on the upper diagonal are the 95% quantile ellipses of the fitted clusters, with color intensity proportional to the mixture probabilities.

Table 3.2: Summary of utility measures with the results for the Concrete industry from the 2007 CMF, for the MNAR imputation scenario with higher probabilities for bottom ranked clusters. The results highlighted in bold correspond to the follow-up sample size that minimized the sum of cost and the transformed measure $\bar{\rho}$. *The values of $\bar{\rho}$ (sd) should be multiplied by 10^{-5}.

(a) New mixture model fit to observed data and follow-up sample $\left(D_R \cup D_{F,\delta}^{(\text{bottom})}\right)$

δ	$\bar{\rho}$ $(sd)^*$	$\bar{\tau}$ (sd)	$\bar{\theta}$ (sd)
0	117.986 (12.423)	5.395 (0.396)	0.958 (0.046)
0.25	60.020 (6.602)	3.866 (0.273)	0.691 (0.038)
0.5	**24.628 (4.492)**	**2.401 (0.275)**	**0.422 (0.039)**
0.75	6.260 (1.712)	1.231 (0.184)	0.220 (0.031)
1	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)

(b) New mixture model fit to follow-up sample only $\left(D_{F,\delta}^{(\text{bottom})}\right)$

δ	$\bar{\rho}$ $(sd)^*$	$\bar{\tau}$ (sd)	$\bar{\theta}$ (sd)
0	–	–	–
0.25	12.833 (5.391)	0.642 (0.284)	0.122 (0.063)
0.5	**3.828 (1.896)**	**0.312 (0.136)**	**0.059 (0.026)**
0.75	1.460 (0.661)	0.201 (0.111)	0.037 (0.023)
1	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)

Table 3.3: Summary of utility measures with the results for the Concrete industry from the 2007 CMF, for the MNAR imputation scenario with higher probabilities for top ranked clusters. The results highlighted in bold correspond to the follow-up sample size that minimized the sum of cost and the transformed measure $\bar{\rho}$. *The values of $\bar{\rho}$ (sd) should be multiplied by 10^{-5}.

(a) New mixture model fit to observed data and follow-up sample $\left(D_R \cup D_{F,\delta}^{(\text{top})}\right)$

δ	$\bar{\rho}$ $(sd)^*$		$\bar{\tau}$ (sd)		$\bar{\theta}$ (sd)	
0	411.057	(19.222)	8.732	(0.285)	1.415	(0.049)
0.25	197.266	(12.833)	6.278	(0.269)	1.026	(0.037)
0.5	**80.489**	**(12.273)**	**4.057**	**(0.282)**	**0.668**	**(0.046)**
0.75	16.758	(2.476)	1.910	(0.155)	0.315	(0.026)
1	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)

(b) New mixture model fit to follow-up sample only $\left(D_{F,\delta}^{(\text{top})}\right)$

δ	$\bar{\rho}$ $(sd)^*$		$\bar{\tau}$ (sd)		$\bar{\theta}$ (sd)	
0	–		–		–	
0.25	12.567	(7.297)	0.541	(0.271)	0.088	(0.049)
0.5	4.731	(2.617)	0.311	(0.177)	0.054	(0.030)
0.75	**1.451**	**(0.820)**	**0.171**	**(0.092)**	**0.029**	**(0.017)**
1	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)

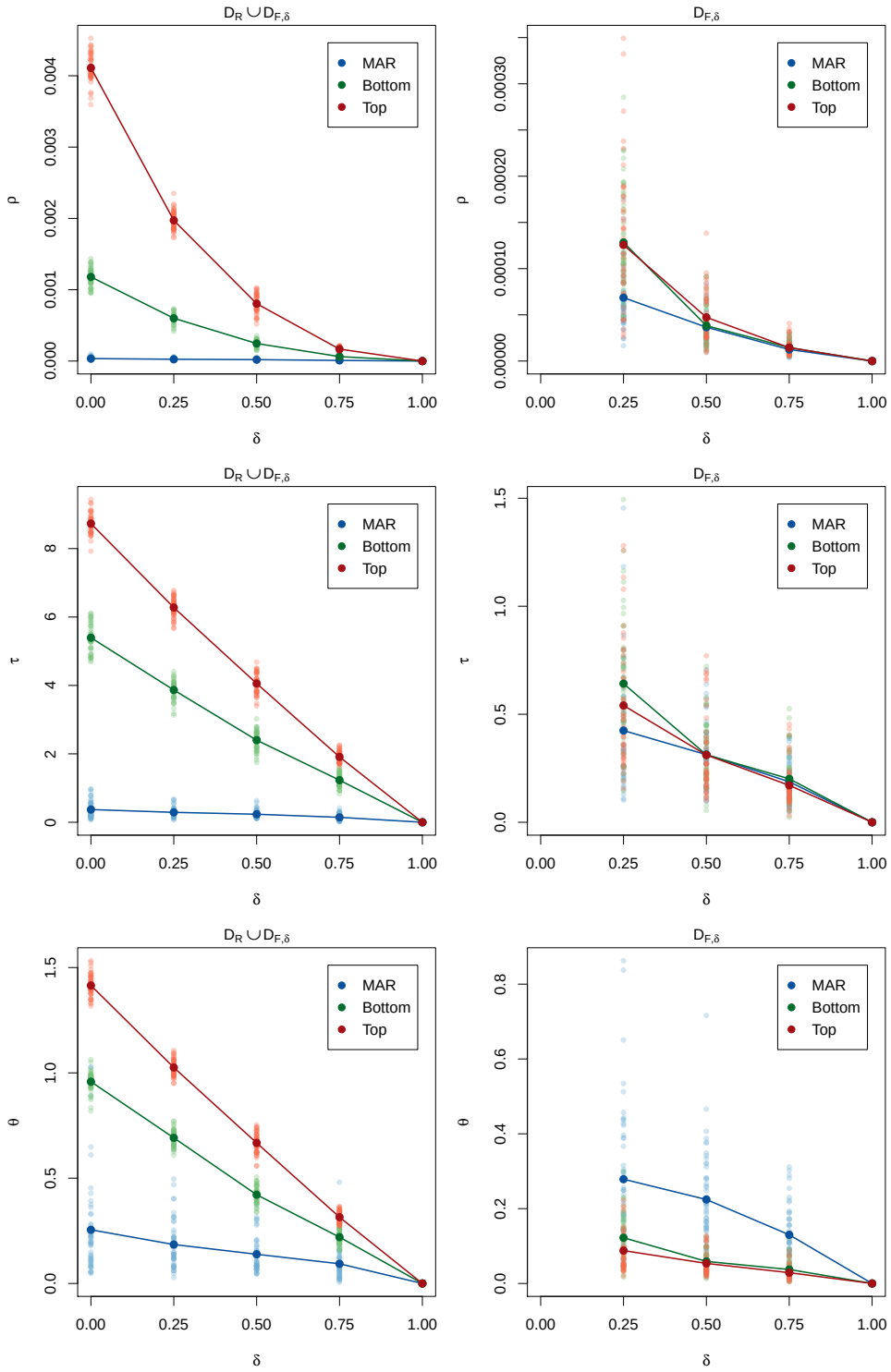


FIGURE 3.6: Summary of utility measures for the three scenarios considered for the Concrete industry from the 2007 CMF. The results for ρ, τ and θ are on the first, second and last row, respectively. The plots on the right column contain the results with the model fit to observed data and follow-up sample, while the plots on the right contain the results with the model fit to follow-up sample only. The faded points are the individual values for each multiple imputation.

We analyze now the results of the different scenarios for the plastic industry. First, we have the pairwise scatterplots of the imputation results under MAR in Figure 3.7 also with the axes removed. For this industry, the mixture model resulted in 8 clusters, with most of the weight concentrated in almost half of them. As in to the concrete industry, we also considered the scenario with higher probabilities for the bottom ranked clusters, which can be seen in Figure 3.8, and the scenario with higher probabilities for the top ranked clusters, which can be seen in Figure 3.9.

The utility measures are summarized in Table 3.4 for the MAR scenario; in Table 3.5 for the MNAR scenario with higher probabilities for the bottom ranked clusters; and in Table 3.6 for the MNAR scenario with higher probabilities for the top ranked clusters. We can also compare the measures for the different scenarios with the plots in Figure 3.10. Again, the values for the MAR cases are smaller than the values of the MNAR cases, since the distributions should be similar. We can also see that the second approach with the model fit with just the follow-up sample yields better results for the MNAR scenarios. If the data are MAR, collecting more data does not seem to improve the utility measures as much as in the MNAR scenarios. When the data are MNAR, there is some improvement in collecting follow-up samples, but it might not be worth the cost as δ gets bigger than 0.5.

The results for both industries confirm that when the data are missing at random, the best results are obtained with the observed data set and follow-up sample. If the nonrespondents are missing not at random, it is better to use the second proposed approach: if collecting a follow-up sample, use this to fit a new mixture model and impute the remaining data. In the MNAR cases, we can see from the tables and plots for both industries that collecting a follow-up sample of 25%-50% of the nonrespondents, can result in utility measures similar to the MAR case. Comparing these values and considering the linear cost increase alongside with the increase in the follow-up sample size, the agency can make their decision about how to proceed

Table 3.4: Summary of utility measures with the results for the Plastic industry from the 2007 CMF, for the MAR imputation scenario. The results highlighted in bold correspond to the follow-up sample size that minimized the sum of cost and the transformed measure $\bar{\rho}$. *The values of $\bar{\rho}$ (sd) should be multiplied by 10^{-5}.

(a) New mixture model fit to observed data and follow-up sample $\left(D_R \cup D_{F,\delta}^{(\text{MAR})}\right)$

δ	$\bar{\rho}$ $(sd)^*$	$\bar{\tau}$ (sd)	$\bar{\theta}$ (sd)
0	8.594 (5.057)	0.402 (0.186)	0.300 (0.175)
0.25	**5.617 (2.175)**	**0.284 (0.143)**	**0.219 (0.139)**
0.5	3.661 (1.910)	0.248 (0.133)	0.199 (0.124)
0.75	1.655 (0.901)	0.154 (0.070)	0.112 (0.066)
1	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)

(b) New mixture model fit to follow-up sample only $\left(D_{F,\delta}^{(\text{MAR})}\right)$

δ	$\bar{\rho}$ $(sd)^*$	$\bar{\tau}$ (sd)	$\bar{\theta}$ (sd)
0	–	–	–
0.25	15.486 (7.486)	0.617 (0.336)	0.430 (0.218)
0.5	**5.674 (2.721)**	**0.340 (0.165)**	**0.252 (0.143)**
0.75	1.825 (0.775)	0.182 (0.119)	0.138 (0.110)
1	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)

with data collection.

As in the concrete industry, we also highlight the results with the values of δ in Table 3.4, Table 3.5 and Table 3.6 that minimize the measure the sum of cost and the transformed measure $\bar{\rho}$. Based on this example, in almost all cases, collecting a follow-up sample with 50% of the nonrespondents gives the best option in the trade-off between cost and utility.

Table 3.5: Summary of utility measures with the results for the Plastic industry from the 2007 CMF, for the MNAR imputation scenario with higher probabilities for bottom ranked clusters. The results highlighted in bold correspond to the follow-up sample size that minimized the sum of cost and the transformed measure $\bar{\rho}$. *The values of $\bar{\rho}$ (sd) should be multiplied by 10^{-5}.

(a) New mixture model fit to observed data and follow-up sample $\left(D_R \cup D_{F,\delta}^{(\text{bottom})}\right)$

δ	$\bar{\rho}$ $(sd)^*$		$\bar{\tau}$ (sd)		$\bar{\theta}$ (sd)	
0	348.335	(27.074)	7.700	(0.402)	0.786	(0.029)
0.25	164.508	(21.269)	5.409	(0.382)	0.558	(0.030)
0.5	**62.564**	**(10.560)**	**3.417**	**(0.318)**	**0.355**	**(0.028)**
0.75	15.779	(2.842)	1.695	(0.204)	0.178	(0.020)
1	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)

(b) New mixture model fit to follow-up sample only $\left(D_{F,\delta}^{(\text{bottom})}\right)$

δ	$\bar{\rho}$ $(sd)^*$		$\bar{\tau}$ (sd)		$\bar{\theta}$ (sd)	
0	–		–		–	
0.25	26.730	(11.243)	0.726	(0.476)	0.077	(0.049)
0.5	**7.122**	**(4.451)**	**0.369**	**(0.286)**	**0.039**	**(0.030)**
0.75	2.906	(1.492)	0.285	(0.185)	0.031	(0.020)
1	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)

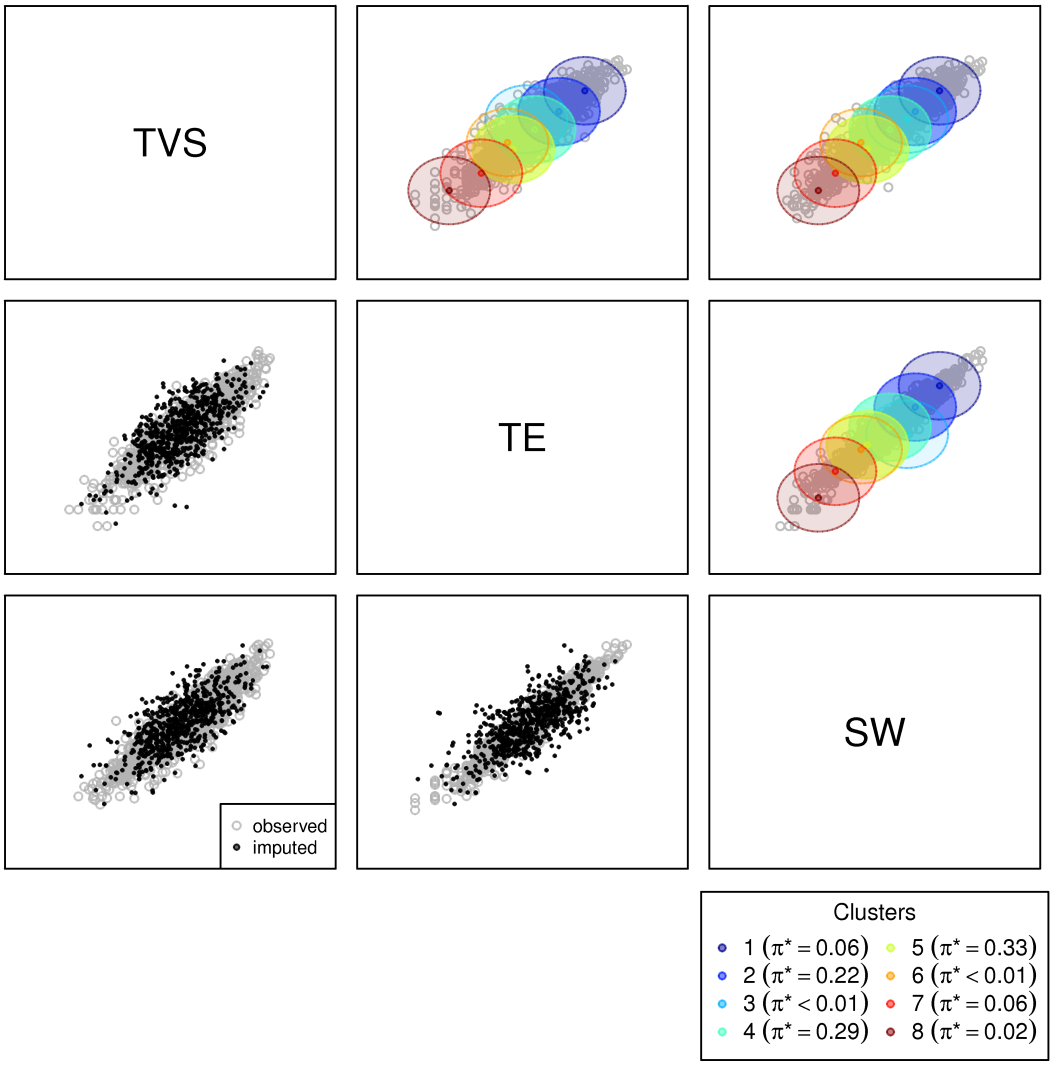


FIGURE 3.7: Pairwise scatterplots with the results for the Plastic industry from the 2007 CMF, for the MAR imputation scenario. Observed points are plotted as gray hollow circles. The black filled circles on the lower diagonal are the imputed points. The colored circles on the upper diagonal are the 95% quantile ellipses of the fitted clusters, with color intensity proportional to the mixture probabilities.

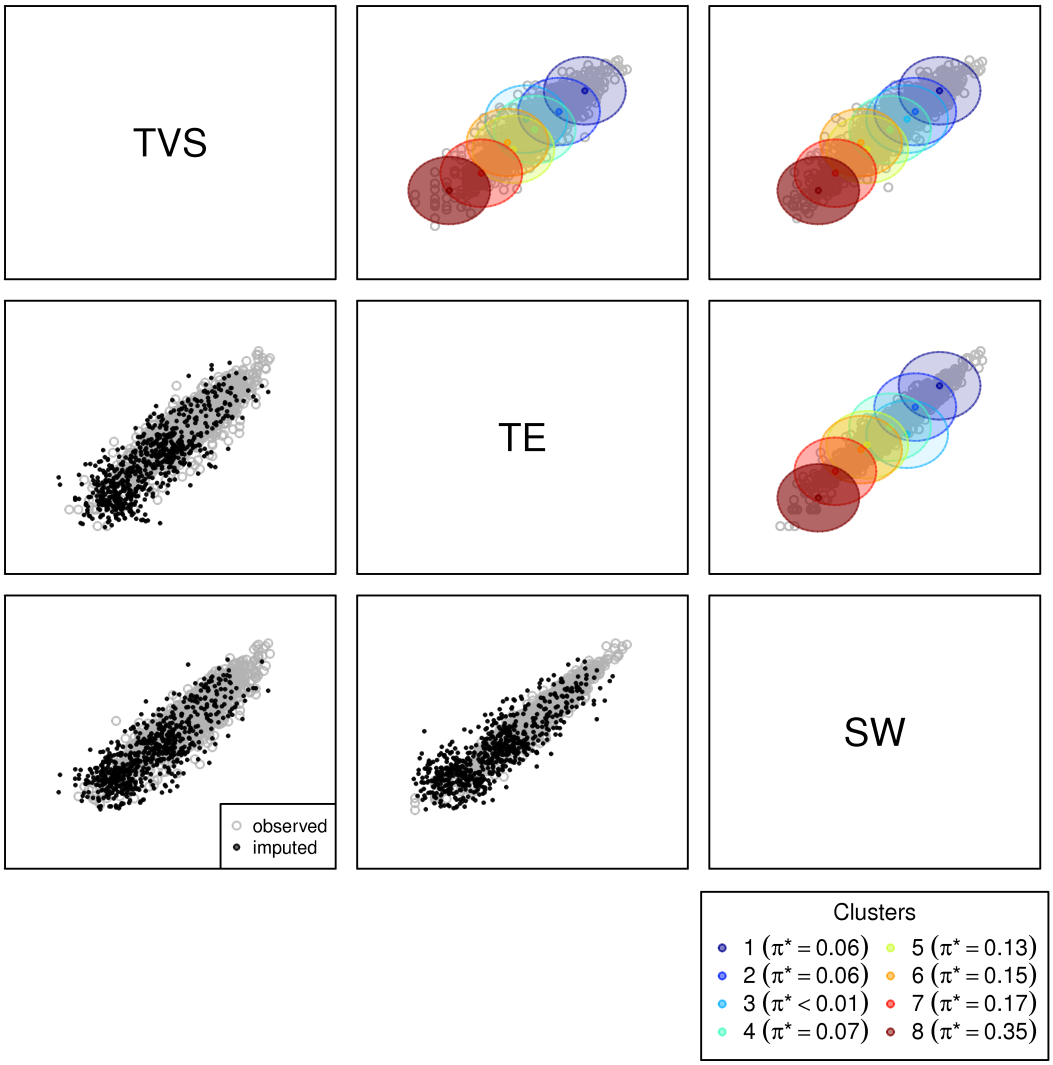


FIGURE 3.8: Pairwise scatterplots with the results for the Plastic industry from the 2007 CMF, for the MNAR imputation scenario with higher probabilities for bottom ranked clusters. Observed points are plotted as gray hollow circles. The black filled circles on the lower diagonal are the imputed points. The colored circles on the upper diagonal are the 95% quantile ellipses of the fitted clusters, with color intensity proportional to the mixture probabilities.

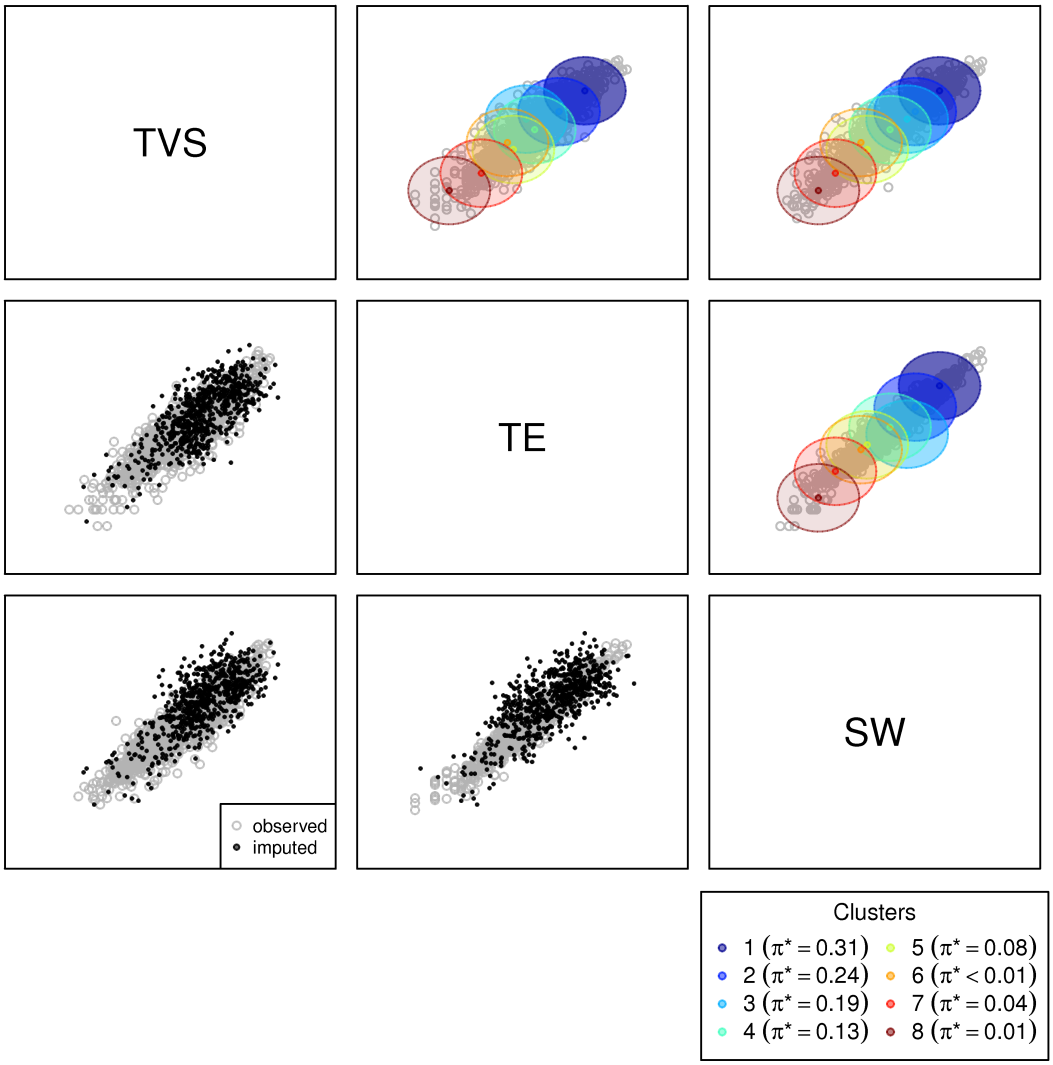


FIGURE 3.9: Pairwise scatterplots with the results for the Plastic industry from the 2007 CMF, for the MNAR imputation scenario with higher probabilities for top ranked clusters. Observed points are plotted as gray hollow circles. The black filled circles on the lower diagonal are the imputed points. The colored circles on the upper diagonal are the 95% quantile ellipses of the fitted clusters, with color intensity proportional to the mixture probabilities.

Table 3.6: Summary of utility measures with the results for the Plastic industry from the 2007 CMF, for the MNAR imputation scenario with higher probabilities for top ranked clusters. The results highlighted in bold correspond to the follow-up sample size that minimized the sum of cost and the transformed measure $\bar{\rho}$. *The values of $\bar{\rho}$ (sd) should be multiplied by 10^{-5}.

(a) New mixture model fit to observed data and follow-up sample $\left(D_R \cup D_{F,\delta}^{(\text{top})}\right)$

δ	$\bar{\rho}$ $(sd)^*$	$\bar{\tau}$ (sd)	$\bar{\theta}$ (sd)
0	170.522 (22.869)	5.295 (0.382)	16.087 (18.513)
0.25	88.387 (13.096)	3.793 (0.380)	11.943 (14.377)
0.5	**35.737 (6.114)**	**2.437 (0.224)**	**7.625 (8.917)**
0.75	8.578 (2.141)	1.141 (0.190)	3.642 (4.301)
1	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)

(b) New mixture model fit to follow-up sample only $\left(D_{F,\delta}^{(\text{top})}\right)$

δ	$\bar{\rho}$ $(sd)^*$	$\bar{\tau}$ (sd)	$\bar{\theta}$ (sd)
0	–	–	–
0.25	17.918 (8.007)	0.623 (0.323)	2.133 (2.956)
0.5	**6.176 (3.219)**	**0.297 (0.181)**	**1.136 (2.212)**
0.75	2.252 (1.253)	0.234 (0.176)	0.730 (0.947)
1	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)

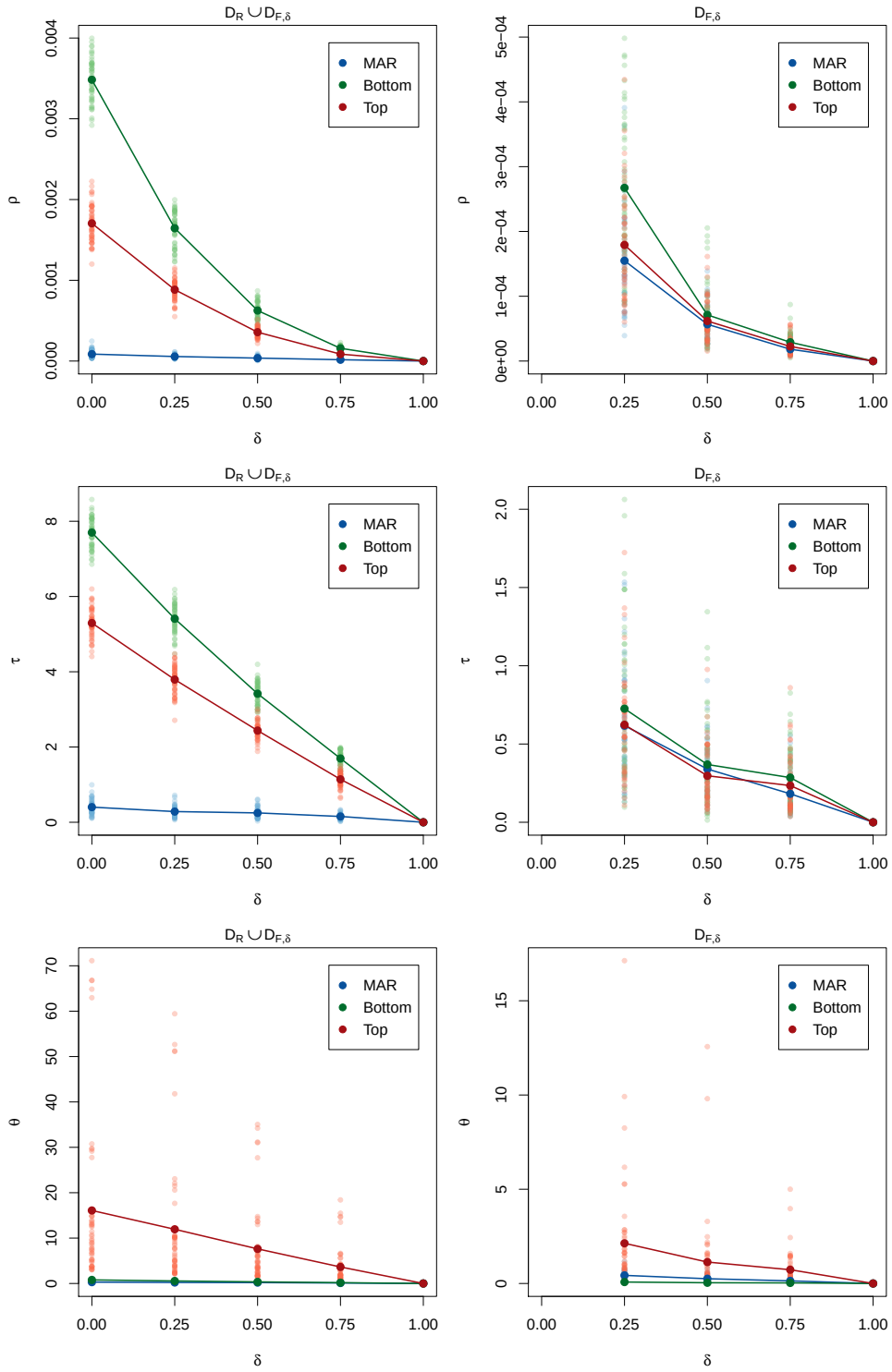


FIGURE 3.10: Summary of utility measures for the three scenarios considered for the Plastic industry from the 2007 CMF. The results for ρ, τ and θ are on the first, second and last row, respectively. The plots on the right column contain the results with the model fit to observed data and follow-up sample, while the plots on the right contain the results with the model fit to follow-up sample only. The faded points are the individual values for each multiple imputation.

3.4 Conclusions

We present an approach for adaptive survey designs using methods for imputation of multivariate continuous data with nonignorable missingness. During an ongoing survey, the agency can decide to stop data collection or obtain follow-up samples based on the information that has been collected so far. This decision is based on the representativeness of the respondents, and how the nonrespondents can affect the inference results. In the case of interrupting data collection and imputing the missing data, we can reduce costs and allocate resources more efficiently.

This adaptive design approach uses the imputation method proposed in Chapter 2, which estimates the observed data distribution with a Dirichlet process mixture of multivariate normals. The model is flexible to capture many distributional features of the observed data, and can be easily modified to reflect different distributions for imputation of the nonrespondents in case of nonignorable missingness. With the development of the NIMC application, the user can easily generate several scenarios to be compared through sensitivity analysis.

Our method is appropriate for sensitivity analysis, and provides a framework to evaluate the impact of extreme scenarios on the results of the imputation. The NIMC application facilitates the process of creating different imputation scenarios, which is an important part of the adaptive design method that we developed. By comparing the scenarios, the user can evaluate the costs and benefits of collecting more data. These benefits can be quantified by the utility measures that we propose, based on propensity scores and marginal statistics of the variables of interest, for different sample sizes.

We demonstrate how to apply the imputation method and obtain the utility measures with data from some industries from the 2007 U.S. Census of Manufactures. The results show how some assumptions about the distribution of the nonrespondents

can be implemented to generate more representative imputed data. The application to the CMF data also demonstrates how much the inference results can be affected by different MNAR scenarios, and how much these impacts are reduced by collecting more waves of data.

If the decision is to stop data collection after considering the measures under different scenarios, the nonrespondents have to be imputed for release of a completed data set. In that case, the user might consider using external information or even a different imputation model. For example, in the CMF there are some data available from administrative records about employment, sales and payroll, collected from other sources, such as the Internal Revenue Service (IRS). These administrative records are highly correlated to the actual recorded responses. Thus, they could be modeled jointly with the variables of interest. The missing data can thengraybe imputed conditioned on these other variables, similarly to the item nonresponse imputation that we described in Section 3.2.2. Here, we assume that given other variables, the missing items are missing at random. A natural extension of this method is to enable imputation for nonignorable item nonresponse.

Another direction for future work is to consider nonresponse in the follow-up sample, which would arise in practice. This problem can be handled by either considering that the second wave nonrespondents are missing at random from the first wave nonrespondents group, or that they are missing not at random. In the first case, they can be imputed together with the units not in the follow-up sample. In the second case, we can fit a new model to the observed part of the follow-up sample, and impute the nonrespondents from a different distribution by changing the mixture probabilities again.

4

Imputation of confidential data sets with spatial locations using disease mapping models

This chapter presents an approach to protect confidentiality in public use data with geographic identifiers. The presentation closely follows the work of Paiva et al. (2014).

4.1 Introduction

Many government agencies, research centers, and principal investigators collect data that they intend to disseminate broadly. These organizations, henceforth all called agencies, often are ethically and even legally obligated to protect data subjects' identities and sensitive attributes. This is particularly challenging when agencies seek to include fine-scale geographic variables on the files. For example, including exact addresses could enable ill-intentioned users to match names to addresses using public records, thereby revealing data subjects' identities. Even modestly coarse geography, like street block or census tract of residence, can be risky when demographic or other readily available attributes are on the file, which when combined may result in

identifications.

To reduce disclosures, most agencies aggregate geographies to high levels before sharing data, if they share geography at all (National Research Council, 2007). For example, agencies following the safe harbor provisions of the U.S. Health Insurance Portability and Accountability Act (HIPAA), which regulates sharing of personal health information, are required to release geographic units comprising at least 20,000 people. As another example, the U.S. Census Bureau does not release geographic identifiers below aggregates of at least 100,000 people in public use files of census data. While such aggregation preserves analyses at the level of aggregation, it can disable small area estimation, mask local spatial dependencies, and create ecological inference fallacies at lower levels of aggregation. Other strategies for protecting geography include adding random noise to locations, e.g., Armstrong et al. (1999), VanWey et al. (2005); or swapping individuals across locations, e.g., Zayatz (2007), Young et al. (2009).

An alternative framework for protecting geographies was proposed by Wang and Reiter (2012): replace actual locations with locations simulated from statistical models. Specifically, Wang and Reiter treat the precise latitude and longitude of each location as a bivariate outcome to be predicted from the other attributes on the file. After fitting a prediction model—regression trees in their illustrative example—they generate new, replacement locations for each individual on the file. To account for the uncertainty introduced by simulation and thereby enable estimation of variances, they recommend that agencies generate $m > 1$ versions of the data sets for dissemination. Such data sets can protect confidentiality, since identification of units and their sensitive data can be difficult when the geographies in the released data are not actual, collected values. And, when the simulation models faithfully reflect the relationships in the collected data, the shared data can preserve spatial associations, avoid ecological inference problems, and facilitate small area estimation. A related

approach was used by Machanavajjhala et al. (2008), who use multinomial regressions to synthesize the street blocks where people live conditional on the street blocks where they work and other block-level attributes.

The approach in Wang and Reiter (2012) requires that the agency knows the latitude and longitude of each location. These may not be available, at least not immediately and without additional cost for geocoding. Further, in many settings, the spatial distribution of attributes can be multi-modal and complex, so that it is difficult to identify good-fitting bivariate regression approaches. Motivated by these limitations, and with a goal of accurately modeling the spatial distribution of locations, we propose to use areal level spatial models, often referred to as disease mapping models (Clayton and Kaldor, 1987; Besag et al., 1991; Clayton and Bernardinelli, 1992; Wakefield, 2007), as engines for generating simulated locations. The basic idea is to (i) tile the spatial surface in ways intended to ensure adequate confidentiality protection, (ii) estimate disease mapping models that predict observed, areal-level counts from attributes on the file, and (iii) use the estimated models to sample multiple, new locations for each individual based on its attribute pattern. This approach applies most naturally for areal geographies like census tracts or street blocks, but it also can be applied with finer-grain coordinates like point locations after an initial aggregation.

We focus exclusively on methods for altering geography, leaving attributes at their original values. We note, however, that agencies might decide instead or in addition to alter the attributes on the file to strengthen the confidentiality protection (Willenborg and de Waal, 2001; National Research Council, 2005; An et al., 2010). As examples, Zhou et al. (2010) use spatial smoothing to mask non-geographic attributes in a Medicare database, leaving original locations unperturbed; and, the Census Bureau swaps the attribute data for individuals in neighboring areas when creating the public use microdata files for the decennial census. Such methods could be

applied after the generation of synthetic geographies; see Wang and Reiter (2012) for further discussion.

The remainder of the chapter is organized as follows. In Section 4.2, we present the areal spatial modeling approach for generating synthetic geography. In Section 4.3, we describe several metrics for assessing the disclosure risks in the released synthetic data sets. We also review how one obtains point and interval estimates from such data sets. In Section 4.4, we illustrate the approach by generating multiply-imputed, partially synthetic versions of a spatially-referenced data set describing causes of death in Durham, North Carolina. Finally, in Section 4.5, we conclude with discussion of implementation of the approach.

4.2 Areal Spatial Models for Data Synthesis

To provide context for the approach, we introduce a scenario that motivated our investigations. Suppose a state public health agency seeks to release counts of lung cancer incidence by sex, race, and age (categorized) for each street block in the state. The agency owns the appropriate data, but it cannot release the exact counts in the blocks because of confidentiality promises.

More formally, let $D = (S, X)$ comprise data on n individuals, where $S = (s_1, \ldots, s_n)$ includes each individual's location and X is the $n \times p$ matrix of each individual's non-spatial attributes. As in the motivating scenario, let the p attributes $(X_1, \ldots, X_p)$ be discrete-valued. For $k = 1, \ldots, p$, let d_k be the number of levels in X_k. Without loss of generality, assume the values in each $X_k = (1, \ldots, d_k)$. Let $b = 1, \ldots, B$ index each distinct attribute pattern in X, where $B \leq \prod_{k=1}^{p} d_k$. For each (b, k), let $x_k^{(b)}$ be the value of X_k in b.

We seek to model where people with certain demographic patterns are likely to reside. To do so, we assume that the area of interest can be divided into a grid

comprising G cells. The grid cells may comprise pre-existing areal units, such as collections of census tracts or street blocks. Alternatively, for point-resolved geography, they may be imposed by the agency for reasons related to computational convenience and, as we shall discuss in Section 4.3, reduction of confidentiality disclosure risks. Let each grid cell be indexed with $i = 1, \ldots, G$. For each (i, b), let $c_i^{(b)}$ be the number of observations in cell i with attributes b. For $k = 1, \ldots, p$, let $Z_k^{(b)}$ be a $d_k \times 1$ vector comprising a one at position $x_k^{(b)}$ and zeros elsewhere. We propose to estimate spatial models of the form,

$$\begin{aligned} c_i^{(b)} &\sim Poisson(\lambda_i^{(b)}) \\ \log \lambda_i^{(b)} &= \mu + \sum_{k=1}^{p} \alpha_k' Z_k^{(b)} + \theta_i + \sum_{k=1}^{p} \phi_{ik}' Z_k^{(b)} + \epsilon_i^{(b)}. \end{aligned} \tag{4.1}$$

Here, μ is the overall intercept; each $\alpha_k = (\alpha_{k1}, \ldots, \alpha_{kd_k})$ is a $d_k \times 1$ vector of main effects for attribute k; θ_i is a grid-specific spatial effect; and each $\phi_{ik} = (\phi_{ik1}, \ldots, \phi_{ikd_k})$ is a $d_k \times 1$ vector of grid-specific spatial effects for attribute k. For identifiability, we set each α_{k1} and each ϕ_{ik1} equal to zero. We note that one also can include interactions among the attributes, as well as flexible functions of (grid-level) continuous attributes. The spatial effects allow $\lambda_i^{(b)}$ to vary by grid cell and attribute pattern. The $\epsilon_i^{(b)}$ is an error term that allows for additional flexibility in the modeling. The model implies that, within any grid cell, the spatial intensities are assumed homogeneous over all points in that cell; in Section 4.5 we suggest related approaches based on point process modeling that do not make this assumption. We note that the model in (4.1) is akin to the areal-level spatial ANOVA model in Kaufman and Sain (2010).

To induce spatial correlation among neighboring grid cells, we use the intrinsic CAR model (Banerjee et al., 2004) as the prior distribution for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_G)$.

Specifically, we assume for all i that

$$\theta_i|\boldsymbol{\theta}_{-i} \sim N(\bar{\theta}_i, \sigma_\theta^2/n_i). \tag{4.2}$$

Here, $\boldsymbol{\theta}_{-i}$ includes the values of θ_j for all $j \neq i$, and $\bar{\theta}_i$ is the average of the n_i values of θ_j for cells j that are neighbors of cell i. We define neighbors to be grid cells that share vertices. Using an analogous notation, for $\{ikj : i = 1, \ldots, G, k = 1, \ldots, p, j = 2, \ldots, d_k\}$, we assume

$$\phi_{ikj}|\boldsymbol{\phi}_{-i,kj} \sim N(\bar{\phi}_{ikj}, \sigma^2_{\phi_{kj}}/n_i). \tag{4.3}$$

Following Banerjee et al. (2004), to ensure identifiability we constrain the elements of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}_{kj}$ so that $\sum_{i=1}^G \theta_i = 0$ and $\sum_{i=1}^G \phi_{ikj} = 0$ for all (kj).

We include $\epsilon_i^{(b)}$ to capture residual variation in the Poisson rates that is not explained by the covariates, thereby adding flexibility to the modeling. The $\epsilon_i^{(b)}$ also can account partially for (unspecified) non-linearities in the predictor function for the Poisson rates. We assume that

$$\epsilon_i^{(b)} \sim N(0, \sigma_\epsilon^2). \tag{4.4}$$

For prior distributions on non-zero hyperparameters, we use

$$\begin{aligned}
\mu &\sim N(0, v_\mu) && (4.5)\\
\alpha_{kj} &\sim N(0, v_{\alpha_k}) \text{ for all } (kj) && (4.6)\\
1/\sigma_\theta^2 &\sim \text{Gamma}(a_\theta, b_\theta) && (4.7)\\
1/\sigma^2_{\phi_{kj}} &\sim \text{Gamma}(a_{\phi_k}, b_{\phi_k}) \text{ for all } (kj) && (4.8)\\
1/\sigma_\epsilon^2 &\sim \text{Gamma}(a_\epsilon, b_\epsilon). && (4.9)
\end{aligned}$$

Here, we recommend setting all variances $v_{\ldots}$ to large values, e.g., around 100, to represent vague prior information. We recommend typical vague prior specifications

for $1/\sigma^2_\theta$ and each $1/\sigma^2_{\phi_k}$, for example setting $(a_\theta = b_\theta = a_{\phi_k} = b_{\phi_k}) = .1$ for all (k, b). In the application described in Section 4.4.1, results were not sensitive to other common hyperparameter choices (in particular, $a_\theta = a_{\phi_k} = 2$ and $b_\theta = b_{\phi_k} = 1$). For $1/\sigma^2_\epsilon$, we recommend following the suggestions of Banerjee et al. (2004, eq. 5.48) when specifying (a_ϵ, b_ϵ). They advise to account for the difference in the precision dimensions between the pure error term and the CAR model when setting these hyperparameters. We present an example of this specification in Section 4.4.

Posterior distributions of the parameters of this model, and hence of the λs, can be estimated via Markov Chain Monte Carlo algorithms. In the applications in Section 4.4, we sample from full conditionals not in closed form using adaptive rejection sampling (Gilks and Wild, 1992).

After estimating the posterior distributions of $\boldsymbol{\lambda} = \{\lambda_i^{(b)}\}$, the agency can generate synthetic locations for the n individuals on the file. To begin, the agency takes a single draw of $\boldsymbol{\lambda}$, say $\boldsymbol{\lambda}^{(l)}$, from its posterior distribution. For all (i, b), the agency computes

$$p_i^{(lb)} = \lambda_i^{(lb)} / \sum_{i=1}^{G} \lambda_i^{(lb)}. \tag{4.10}$$

For each individual with attribute pattern b, the agency randomly and independently samples its grid cell according to the probabilities in $(p_1^{(lb)}, \ldots, p_G^{(lb)})$. When convenient, the sampled grid cells can serve as one set of synthetic locations. Alternatively, the agency can sample finer coordinates from inside the grid cell, for example sampling uniformly from feasible geographic locations (e.g., capable of being residences) inside the cell. The result is one set of synthetic locations, $\tilde{S}^{(l)} = (\tilde{s}_1^{(l)}, \ldots, \tilde{s}_n^{(l)})$, which when attached to X results in one partially synthetic data set, $\tilde{D}^{(l)} = (\tilde{S}^{(l)}, X)$. The agency repeats the process independently m times to obtain m sets of synthetic locations, $\tilde{S} = (\tilde{S}^{(1)}, \ldots, \tilde{S}^{(m)})$, and corresponding data sets, $\tilde{D} = (\tilde{D}^{(1)}, \ldots, \tilde{D}^{(m)})$,

which are released to the public. In practice, to obtain approximately independent realizations of the synthesis process, the agency either (i) can run m MCMC chains initiated at dispersed starting values and use the final draw of $\boldsymbol{\lambda}$ for each chain, or (ii) run one long MCMC chain thinned so that autocorrelations among estimated components of $\boldsymbol{\lambda}$ are approximately zero.

It is worth noting that two records close in space in the original data will not necessarily be close in space in the synthetic data, since their synthetic locations (grid cells and coordinates) are independently generated from the estimated Poisson models. Arguably, such possible movement is necessary to reduce disclosure risks sufficiently.

A key feature of this modeling is the choice of the grid size. Intuitively, a very thin grid (with a large number of grid cells) allows for greater heterogeneity in the tiled rate surface. When such heterogeneity is an important feature in the data, this should allow the pattern of synthetic geographies to mimic the observed pattern more faithfully. However, a very thin grid also could result in synthetic locations that are too close to the original ones, which could fail to protect confidentiality. Conversely, a very coarse grid tends to improve protection at cost of reduced data quality. This suggests that agencies can benefit from examining trade offs in disclosure risk and data quality for multiple candidate grid sizes. This requires quantifying disclosure risks and data quality, which we now discuss.

4.3 Disclosure Risk and Data Utility

In this section, we describe an approach to assessing disclosure risks of the partially synthetic data with simulated geographies. We focus on computing the probabilities that individuals' true areal geographies can be learned from the released data, building on ideas developed by Duncan and Lambert (1989) and applied subsequently by several authors (e.g., Fienberg et al. (1997); Reiter (2005); Drechsler and Reiter

(2008); Reiter and Mitra (2009)). We use functions of these probabilities to create risk metrics for both attribute disclosure (an intruder learns the value of true areal geography) and identification disclosure (an intruder learns the identity of some record).

We also review the inferential methods for partially synthetic data (Reiter, 2003). These methods enable comparisons of inferences based on the original and synthetic data, which represent our primary approach to evaluate the utility of the synthetic data.

4.3.1 Risk Measures

To learn geographies, we assume that the intruder utilizes all information at her or his disposal. This includes information released about the synthetic data model, which we denote with M. For example, M could include mathematical descriptions corresponding to (4.1) – (4.9), including the definitions of the grid cells. Alternatively, M could include the code used to fit the models without parameter estimates (releasing parameter estimates could leak too much information about S). The intruder also may possess auxiliary information about the geographies of the records on the file, which we denote with A. For example, A could include the geographies of some subset of individuals on the file, or it could be empty.

Using this information, the intruder seeks to determine the probable values of s_t for some record t. We assume the intruder knows that t is in the sample (but does not know its location). This assumption can be relaxed; see Drechsler and Reiter (2008) for a general strategy to do so. The intruder need not affiliate a particular row in $\tilde{D}$ with t to determine probable values of s_t; indeed, for patterns b such that $\sum_i c_i^{(b)} > 1$ unique affiliation is impossible. Thus, for any particular t and potential

location s, the intruder seeks to estimate

$$\rho_t^s = P(s_t = s|\tilde{D}, A, M) = cP(\tilde{S}|s_t = s, X, A, M)P(s_t = s|X, A, M)$$
$$= c\left(\int P(\tilde{S}|s_t = s, X, A, M, \boldsymbol{\lambda})P(\boldsymbol{\lambda}|s_t = s, X, A, M)d\boldsymbol{\lambda}\right) P(s_t = s|X, A, M) \tag{4.11}$$

over all feasible s, where c is a normalizing constant. For any record with a unique attribute pattern, i.e., a b such that $\sum_i c_i^{(b)} = 1$, $\{\rho_t^s\}$ represents the posterior distribution of s_t for that particular record. The interpretation of ρ_t^s is more subtle for records with b such that $\sum_i c_i^{(b)} > 1$ and depends on the nature of A. For example, if an intruder knows the locations of all records in the sample with attribute pattern b except t, then $\{\rho_t^s\}$ again represents the posterior distribution of s_t for one particular record. With other forms of A, intruders may be able to interpret $\{\rho_t^s\}$ only as a distribution for all records with attribute pattern b.

We assume that the intruder selects the s yielding the maximum ρ_t^s as a best guess for s_t. Arguably the most that an intruder can learn from $\tilde{D}$ (beyond X) is the grid cell to which the individual belongs, since finer-grain locations within any cell are randomly sampled within the cell. Hence, we suppose the intruder's goal in computing (4.11) is to find the correct grid cell. Thus, we let S and $\tilde{S}$ in (4.11) comprise grid cells.

Conceptually, $P(s_t = s|X, A, M)$ represents the intruder's prior beliefs about the grid cell of individual t, and $\tilde{S}$ serves to sharpen those beliefs. Effectively, the intruder takes guesses at the true grid cell of individual t according to the prior beliefs. Guesses that result in relatively low probability of generating $\tilde{S}$ (given X, A, and M) are downweighted compared to guesses that result in relatively high probability of generating $\tilde{S}$. By evaluating the probabilities of generating $\tilde{S}$ over all possible s, the intruder determines the *a posteriori* best estimate.

Of course, it is impossible for agencies to know any particular intruder's prior

beliefs. Instead, agencies can adopt the recommendation of Skinner (2012) and evaluate risks under reasonable prior distributions. For example, the agency can use a uniform distribution over all grid cells in the population that include individuals with the same attribute pattern b as individual t. This reflects vague prior knowledge about s_t. In the absence of population counts by attribute patterns per grid cell, the agency can allow the support to include the entire grid.

Similarly, it is impossible for the agency to know the auxiliary information possessed by intruders. One approach, which we adopt here, is to evaluate risks under a "worst case" scenario by assuming that the intruder knows the grid cells of all individuals except one particular t, i.e., the intruder knows $s_{t'}$ for all $t' \neq t$. Call this set S_{-t}. In addition to offering risk estimates for intruders with very strong prior knowledge, setting $A = S_{-t}$ greatly facilitates computation as we describe in Section 4.3.1.

Computational Methods

The form of (4.11) when $A = S_{-t}$ suggests a Monte Carlo approach to estimation of ρ_t^s. First, for any proposed value s, the agency replaces s_t with s to form a new set of locations, $S_t^s = (s_t = s, S_{-t})$, attached to original X. Second, treating (S_t^s, X) as if it were the collected data, the agency samples many values of $\boldsymbol{\lambda}$. Third, for each sampled $\boldsymbol{\lambda}$, the agency computes the probability of generating the released $\tilde{S}$, and averages these probabilities. The agency repeats these three steps for all values of s, which allows computation of the normalizing constant in (4.11) and hence ρ_t^s for all s.

To draw new $\boldsymbol{\lambda}$s for each (S_t^*, X), one approach is to re-estimate the model in (4.1) – (4.9). This is computationally intensive, however, as the the agency needs to estimate G models per t. Instead, we suggest using the sampled values of $\boldsymbol{\lambda}$ from $p(\boldsymbol{\lambda}|D)$ as proposals for an importance sampling algorithm (Robert and Casella,

2005, Chapter 3). As a brief review of importance sampling, suppose we seek to estimate the expectation of some function $g(\boldsymbol{\lambda})$, where $\boldsymbol{\lambda} \sim f(\boldsymbol{\lambda})$. Further suppose that we have available a sample $(\boldsymbol{\lambda}^{(1)}, \ldots, \boldsymbol{\lambda}^{(L)})$ from a convenient distribution $f^*(\boldsymbol{\lambda})$ that differs from $f(\boldsymbol{\lambda})$. We can estimate $E_f(g(\boldsymbol{\lambda}))$ using

$$E_f(g(\boldsymbol{\lambda})) \approx \sum_{j=1}^{L} g(\boldsymbol{\lambda}^{(j)}) \frac{f(\boldsymbol{\lambda}^{(j)})/f^*(\boldsymbol{\lambda}^{(j)})}{\sum_{h=1}^{L} f(\boldsymbol{\lambda}^{(h)})/f^*(\boldsymbol{\lambda}^{(h)})}. \tag{4.12}$$

We note that (4.12) only requires that $f(\boldsymbol{\lambda})$ and $f^*(\boldsymbol{\lambda})$ be known up to normalizing constants.

We implement importance sampling algorithms to approximate the integral in (4.11). For any proposed $s_t = s$, we set $g(\boldsymbol{\lambda}) = cP(\tilde{S}|S_t^s, X, M)$ and seek to approximate its expectation with respect to $f(\boldsymbol{\lambda}) = P(\boldsymbol{\lambda}|S_t^s, X, M)$. To facilitate computation, we work with each set of synthetic locations $\tilde{S}^{(l)}$ separately, since

$$P(\tilde{S}|S_t^s, X, M) = \prod_{l=1}^{m} P(\tilde{S}^{(l)}|S_t^s, X, M). \tag{4.13}$$

Given a sampled value of $\boldsymbol{\lambda}$, we have

$$P(\tilde{S}^{(l)}|S_t^s, X, M, \boldsymbol{\lambda}) = \prod_{i=1}^{G} \prod_{b=1}^{B} \left(p_i^{(lb)}\right)^{\tilde{c}_i^{(lb)}}, \tag{4.14}$$

where $\tilde{c}_i^{(lb)}$ is the count of synthetic points with attribute pattern b in grid cell i from set l, and $p_i^{(lb)}$ is computed as in (4.10) with the sampled $\boldsymbol{\lambda}$. We next set $(\boldsymbol{\lambda}^{(1)}, \ldots, \boldsymbol{\lambda}^{(L)})$ equal to L draws of $\boldsymbol{\lambda}$ already available from the estimated posterior distribution based on D; hence, we set $f^*(\boldsymbol{\lambda}) = f(\boldsymbol{\lambda}|S, X, M)$. Following (4.1) – (4.9), the only differences in the kernels of $f(\boldsymbol{\lambda})$ and $f^*(\boldsymbol{\lambda})$ include (i) the components of the likelihood associated with the counts on the grid cells s and s_t for attribute pattern b and (ii) the normalizing constants for each density. Hence, after computing the

normalized ratio in (4.12), we are left with the expression,

$$P(\tilde{S}^{(l)}|S_t^s, X, M) \approx \sum_{j=1}^{L} \left(\prod_{i=1}^{G} \prod_{b=1}^{B} \left(p_i^{(jb)} \right)^{\tilde{c}_i^{(lb)}} \right) \left(\frac{\lambda_s^{(jb)} / \lambda_{s_t}^{(jb)}}{\sum_{h=1}^{L} \lambda_s^{(hb)} / \lambda_{s_t}^{(hb)}} \right). \tag{4.15}$$

We repeat this computation for $l = 1, \ldots, m$ times, plugging the m results into (4.13). Finally, to approximate ρ_t^s, we compute (4.13) for each s and multiply each resulting value by $P(s_t = s|X, S_{-t}, M)$, and we normalize the collection of G results (hence, computation of c is never required). As a note on computation, the terms in (4.14) for which b does not match the attributes of record t cancel when normalizing, so that one can replace the expression in (4.14) with $\prod_{i=1}^{G} (\lambda_i^{(b)})^{\tilde{c}_i^{(lb)}}$.

Although the importance sampling uses the actual values of S to make proposals for $\boldsymbol{\lambda}$, any S^* could be used. Hence, intruders are able to utilize these approximations as well.

Setting $A = S_{-t}$ simplifies computation immensely, in that when computing ρ_t^s we have to impute new values only for s_t. In contrast, to compute ρ_t^s when $A \neq S_{-t}$, the intruder needs to impute possible values for all unknown geographies. This introduces a potentially large number of computations. One case of particular interest is when A is empty, representing no intruder knowledge. To avoid imputing all of S, one rough approximation is to use each of the m sets of $\tilde{S}_{-t}^{(l)}$ as a draw of S_{-t}, and average the m resulting values of ρ_t^s.

Summary Measures

After obtaining the posterior probabilities, agencies need to summarize these probabilities to evaluate individual and file level disclosure risks. We now present four such risk measures. Each is based on the assumption that the intruder uses the cell s with maximum ρ_t^s as the best guess for s_t.

The first measure assesses the risks that intruders learn true grid cells given the synthetic data; hence, it is an attribute disclosure risk measure. For all $t = 1, \ldots, n$

individuals in the file, let $r_t = 1$ if the maximum posterior probability for record t happens to be on the true s_t (with no ties), and let $r_t = 0$ otherwise. That is, for all t, let $r_t = \mathbf{1}_{\{\arg\max_s(\rho_t^s)=s_t\}}$. A file level risk measure is the percentage of records with $r_t = 1$, i.e.,

$$R_{all} = \sum_{t=1}^{n} r_t/n. \tag{4.16}$$

Intuitively, smaller values of R_{all} are preferable to larger values for confidentiality protection.

As noted by many experts in disclosure estimation (e.g., Skinner and Shlomo (2008)), agencies pay special attention to risks for records with unique combinations of variables in the sample (although arguably uniqueness in the population is more relevant). Singletons are more likely to be identified, since matches to external data are guaranteed to be correct (assuming no errors in matching). With this issue in mind, we introduce a measure that focuses on individuals with unique combinations of (i, b). Formally, for all $t = 1, \ldots, n$, let $c_{(t)}$ be the count of individuals in D matching the grid cell and attribute pattern of t. Let $a_t = 1$ if $c_{(t)} = 1$, and let $a_t = 0$ if $c_{(t)} > 1$. The second risk measure is the percentage of records with unique patterns that the intruder correctly locates,

$$R_{unq} = \frac{(\sum_{t=1}^{n} a_t r_t)}{(\sum_{t=1}^{n} a_t)}. \tag{4.17}$$

Both R_{all} and R_{unq} do not distinguish between intruders whose best guess is close (but not equal) to the actual grid cell and whose best guess is far from the actual grid cell. To distinguish these, we present a third risk measure based on distances. For each t, let d_t be the distance between s_t and the grid cell with the maximum probability, so that

$$d_t = \|s_t - \arg\max_s(\rho_t^s)\|. \tag{4.18}$$

The agency can assess the distributions of d_t over all t to determine if, for example, distances tend not to be concentrated at small values. We compute d_t as the Euclidean distance between s_t and the centroid of the grid cell with maximum probability. Thus, when $r_t = 1$, d_t is bounded by half of the diagonal of a grid cell.

While R_{all}, R_{unq}, and the distribution of d_t summarize risks that intruders learn true geographies, they are not readily interpretable as measures of identification disclosure risk. In particular, in some grid cells many records have the same attribute pattern b, so that intruders cannot distinguish between them. For an extreme example, consider using only one grid cell for the entire area. Here, $r_t = 1$ for all t, since *a priori* everyone is guaranteed to be in the cell. Thus, $R_{all} = R_{unq} = 1$, and all d_t are equal. However, since coordinates are sampled randomly within the single cell, releasing $\tilde{S}$ introduces zero risks that individual records will be identified (assuming the study area is already known to the intruder).

To quantify identification disclosure risk, we use a measure similar to one presented in Reiter (2005). For each t, we compute $z_t = r_t/c_{(t)}$. This corresponds to the probability that an intruder guesses correctly when randomly choosing a match from among the $c_{(t)}$ qualifying records with the same attribute pattern and grid cell as record t. A file level risk measure is

$$R_{id} = \sum_{t=1}^{n} z_t/n. \tag{4.19}$$

Thus, R_{id} can be considered the expected number of correct identifications when randomly choosing a match. Implicit in this interpretation is the assumption that the intruder knows that record t is among the n records collected in the original survey. Agencies can relax this assumption by replacing $c_{(t)}$ with the number of individuals in the population (not D) with the same attribute pattern and grid cell as record t. When this population count is unknown, as is often the case, the agency

must estimate it, for example using log-linear models or other approaches (Skinner and Shlomo, 2008; Forster and Webb, 2007; Manrique-Vallier and Reiter, 2012).

4.3.2 Inferences with partially synthetic data

The inferential methods for partially synthetic data depend on the nature of the analysis, as we now describe. Let Q be a scalar estimand of interest, such as a population mean or regression coefficient. Suppose that, if given D, the analyst would use normal distributions for inference, $(Q - q) \sim N(0, u)$. Here, q is a point estimator of Q such as an unbiased estimator or posterior mean, and u is the associated variance. For each $l = 1, \ldots, m$, let $q^{(l)}$ and $u^{(l)}$ be the estimates of q and u computed with $D^{(l)}$. For inferences about Q the analyst needs the following quantities:

$$\begin{aligned} \bar{q}_m &= \sum_{l=1}^{m} q^{(l)}/m \\ b_m &= \sum_{l=1}^{m} (q^{(l)} - \bar{q}_m)^2/(m-1) \\ \bar{u}_m &= \sum_{l=1}^{m} u^{(l)}/m. \end{aligned}$$

The analyst uses $\bar{q}_m$ as a point estimate of Q with associated variance $T_p = b_m/m + \bar{u}_m$. Inferences are based on t-distributions, $(Q - \bar{q}_m) \sim t_\nu(0, T_p)$ with $\nu_m = (m-1)\left(1 + m\bar{u}_m/b_m\right)^2$ degrees of freedom.

With spatial data, analysts often estimate spatial regression models with Bayesian methods (Banerjee et al., 2004; Gelfand et al., 2006). Here, the combining rules in Reiter (2003) may not apply, particularly when the posterior distributions of parameters are not well-approximated by normal distributions. Instead, analysts can use the approach described by Zhou and Reiter (2010). Specifically, the analyst fits the Bayesian model in each synthetic data set, coming up with m sets of posterior

samples of the parameters of interest. The analyst then mixes all m sets of draws to estimate the posterior distribution.

4.4 Illustrative Application

We now apply the methods of Section 4.2 and Section 4.3 to create and evaluate partially synthetic locations for a subset of North Carolina (NC) mortality records from 2002. Similar data were used by Wang and Reiter (2012). The data include precise longitudes and latitudes of deceased individuals' residences and several variables related to manner of death. As explained by Wang and Reiter (2012), these data are publicly available and so do not require disclosure protection; however, because the data include point-referenced locations that can be revealed for comparisons, they represent an ideal testbed for methods that protect confidentiality of geographies. These data also enable us to illustrate how one can apply areal spatial models to protect confidentiality even with point-referenced data.

We selected a subset of individuals with residences in seven contiguous zip codes in Durham, NC, which is an area of approximately 20 by 20 miles. To mimic the types of variables discussed in the motivating example of Section 4.2, we include as attributes individuals' sex (male or female), race (black or white), age (<60, 60-75, 75-85, or >85 years), education (less than high school, high school, some college, more than 4 years of college) and an indicator if the cause of death was caused by cancer or some failure of the immune system versus all other causes. The final sample includes $n = 6294$ individuals. We include only two races (accounting for 97% of the observations in the full data) for convenience of illustration, as the spatial distribution of these two races is clearly evident in the data; see Figure 4.1(a). Similar exploratory maps indicate that variables other than race are more-or-less randomly distributed across the area.

With this set of variables, analysts might treat the indicator for reason of death,

which we label as Y, as an outcome variable to be predicted from the other variables. Here, the analyst may want to control for spatial dependencies in the prediction model, for example to improve predictive power or to account for heterogeneity not captured by the predictors. However, since Y does not follow strong spatial patterns—see Figure 4.1(b)—it is not a particularly useful outcome for testing how well the synthetic data preserve spatial dependencies. We therefore created a surrogate outcome, $\tilde{Y}$, for which location matters. In particular, for $t = 1, \ldots, n$, we set

$$\tilde{Y}_t \sim \text{Bern}(\pi_t), \quad \text{logit}(\pi_t) = X_t\boldsymbol{\beta} + w(s_t). \tag{4.20}$$

Here, X_t includes main effects for sex, race, age, and education. Each $\beta_k \sim N(0, .3)$, and $w(s_t)$ is a mean-zero Gaussian process (Banerjee et al., 2004) with exponential covariance function such that, for any two locations $s \neq s'$, $\text{Cov}(s, s') = \sigma^2 \exp(-\phi\|s - s'\|)$ with $\phi = .6$ and $\sigma^2 = 1$. As evident in Figure 4.1(c), this results in a stronger spatial pattern than Y. We use $D = (\tilde{Y}, X)$ for all subsequent analyses.

4.4.1 Generation of the synthetic data sets

We generate synthetic locations using the approach in Section 4.2 with all $B = 128$ attribute patterns formed from the variables in $D = (\tilde{Y}, X)$. For illustration, we use square grids with three sets of sizes: 10×10, 20×20, and 30×30. The coordinates were rescaled to fall in $[0, 10] \times [0, 10]$, respecting the original proportion of the horizontal and vertical ranges.

Exploratory data analysis suggests that including only main effects in (4.1) – (4.9) is reasonable for these data. Regardless of grid size, we use the prior distributions in (4.5) – (4.9) with $v_\mu = v_{\alpha_k} = 5$ for all attributes k; with $(a_\theta, b_\theta) = (a_{\phi_k}, b_{\phi_k}) = (.1, .1)$ for all k; and with $(a_\epsilon, b_\epsilon) = ((.1)(.7^2)\bar{n}, .1)$, where $\bar{n}$ is the average number of neighbors per grid cell. These values of (a_ϵ, b_ϵ) are suggested by Banerjee et al.

(2004, equation 5.48) to account for the difference in the dimensions of the spatial and non-spatial effects. We obtained similar results using $(a_\theta, b_\theta) = (a_{\phi_k}, b_{\phi_k}) = (2, 1)$ for all k with $(a_\epsilon, b_\epsilon) = (2, 1/[(.7^2)\bar{n}])$.

We run the MCMC for 10001 iterations, tossing the first 1000 as burn-in. We assess the convergence of the chain by analyzing the trace plots of the main effects of the attributes and the posterior mean surfaces of $\boldsymbol{\lambda}^{(b)}$. Posterior intervals for all coefficients across the three grid sizes are included in Table 4.1. The intervals indicate that race and education are the strongest predictors of the Poisson rates. Figure 4.2 displays the posterior mean surfaces of λ for the 20×20 grid for two attribute patterns. The first corresponds to white women over age 85 with education less than high school and $\tilde{Y} = 0$; this has the highest frequency among the combinations. The second pattern corresponds to black men less than age 60 with more than four years of college and $\tilde{Y} = 1$; this pattern is far less frequent (only eight individuals).

After the burn-in, we sample $m = 10$ synthetic locations following the approach in Section 4.2, using a systematic sample of every one thousandth draw.

Table 4.1: Posterior mean and 95% HPD intervals for the coefficients on the expression for $\log \lambda$, for the different grid sizes

	size 10			size 20			size 30		
	mean	LB	UB	mean	LB	UB	mean	LB	UB
μ	-10.94	-11.25	-10.66	-12.51	-12.87	-12.12	-13.44	-13.80	-13.08
$\alpha_{\tilde{Y}}$	-0.01	-0.15	0.13	-0.10	-0.30	0.10	-0.12	-0.33	0.09
α_{sex}	0.02	-0.09	0.13	-0.02	-0.18	0.14	-0.03	-0.19	0.13
α_{race}	-1.79	-2.12	-1.48	-2.13	-2.61	-1.69	-2.35	-2.88	-1.85
α_{age2}	0.09	-0.11	0.32	0.11	-0.14	0.36	0.11	-0.13	0.37
α_{age3}	0.20	-0.03	0.40	0.18	-0.10	0.45	0.16	-0.13	0.45
α_{age4}	-0.05	-0.27	0.19	-0.13	-0.44	0.17	-0.18	-0.49	0.16
α_{edu2}	-0.08	-0.24	0.08	-0.09	-0.28	0.11	-0.10	-0.30	0.10
α_{edu3}	-0.46	-0.68	-0.25	-0.54	-0.81	-0.27	-0.57	-0.85	-0.29
α_{edu4}	-2.04	-2.48	-1.61	-2.09	-2.60	-1.63	-2.21	-2.74	-1.69

4.4.2 Evaluating the utility of the synthetic data sets

We evaluate the utility of the synthetic data by comparing various analyses on the original data (with $\tilde{Y}$) and synthetic data sets. These analyses include estimates of demographic characteristics by zip code, posterior inference from a spatial regression involving $\tilde{Y}$ on the remaining variables, and maps of synthetic locations by various demographic categories.

Figure 4.3 displays the proportions of black people in each of the seven zip codes for the three grid sizes. Here, we determine the 95% confidence intervals using the methods in Reiter (2003), described in Section 4.3.2. With the 30×30 grid and to a slightly lesser extent the 20×20 grid, the confidence intervals from the synthetic data largely overlap with those based on D. The intervals for the 10×10 grid are not as high quality. Figure 4.4 displays analogous results for the proportion of cases with $\tilde{Y} = 1$. Once again, the intervals based on $\tilde{D}$ and D largely overlap, with generally increasing quality as the grid becomes thinner. For the three different grid sizes, the fraction of total variance (T_p) due to variability in the $m = 10$ synthetic point estimates ($b_m/10$) is typically around 15% (sd=6%). We note that, in both figures, the posterior means for the 30×30 grid occasionally are further from the original proportions than those from other grid sizes; this is due primarily to chance.

To evaluate finer spatial information in the synthetic locations, we next estimate the spatial logistic regression in (4.20) based on $\tilde{D}$. As vague prior distributions, we use

$$\boldsymbol{\beta} \sim N(0, 100\mathbf{I}) \tag{4.21}$$

$$\phi \sim \text{Uniform}(0.6, 2.9) \tag{4.22}$$

$$1/\sigma^2 \sim \text{Gamma}(2, 1). \tag{4.23}$$

The bounds of the prior distribution for the spatial decay parameter ϕ are defined

based on the effective range for distances equal to 10% and 50% of the maximum distance between two points in the data ($\bar{d} = 10.5$). Using the relation between the range and ϕ with the exponential covariance function, we obtain the bounds as approximately $3/(.5\,\bar{d})$ and $3/(.1\,\bar{d})$ (Banerjee et al., 2004).

We estimate the posterior distribution of all parameters using the `spGLM` function from the `spBayes` in R. Since n is relatively large, we fit the model using a predictive process (Banerjee et al., 2008; Finley et al., 2009) with 100 knots to obtain posterior samples of the parameters. We estimate a separate MCMC chain for each synthetic data set, running each for 100000 iterations, and combine the resulting posterior draws to obtain synthetic data inferences. We also estimate the posterior distributions using D for comparison. The variability among the synthetic point estimates contributes on average 2.5% (sd=2%) to the total posterior variances of the parameters. We computed these proportions using b_m/m over the posterior variances.

Figure 4.5 displays the posterior mean and 95% central credible intervals for the coefficients and spatial covariance parameters. The credible intervals for the coefficients based on $\tilde{D}$ are very similar to those based on D. Additionally, the posterior distributions of ϕ based on $\tilde{D}$ are similar to those based on D across all grid sizes. The posterior mean of σ^2 when using the 10×10 grid is noticeably lower than the posterior mean when using D. This gap decreases as we increase the number of grid cells. We believe that the results with 20×20 and 30×30 are close enough to those based on D that many analysts would be comfortable interpreting this spatial regression based on $\tilde{Y}$.

Finally, Figures 4.6 and 4.7 display the distribution of race and $\tilde{Y}$ in the original and four randomly chosen synthetic data sets for the 20×20 grid. Results for the 30×30 grid are similar. The maps confirm the trends from the previous analyses: spatial patterns in these variables are maintained in the synthetic data sets. More detailed evidence of this is evident in Figure 4.8 and Figure 4.9, which display the

original and synthetic points for the two patterns from Figure 4.2: white women over age 85 with education less than high school and $\tilde{Y} = 0$; and black men less than age 60 with more than four years of college and $\tilde{Y} = 1$. Overall, the spatial patterns are approximately preserved, as the points are spread around similar areas. Nonetheless, the synthetic locations still can differ from the original ones, as is necessary to reduce disclosure risks.

4.4.3 Evaluating the risk of the synthetic data sets

For all risk analyses, we adopt the "worst case" scenario assumption outlined in Section 4.3.1, so that for each t we set $A = S_{-t}$. Table 4.2 displays the values of $(R_{all}, R_{unq}, R_{id})$ for the generated $\tilde{D}$ at each grid size. Across all three grids, no more than 9% of all locations have correctly identified grid cells. As evident from R_{unq}, records with unique combinations of (i, b) have slightly higher risks of location disclosure. However, the values of R_{unq} indicate that roughly 90% of cases with unique patterns are not correctly located. The intruder has no way of determining which among these cases are correctly located. The values of R_{all} and R_{unq} are larger for the 10×10 grid than the 20×20 grid. As G gets smaller, the area of individual grid cells increases, so that the intruder has greater chances of guessing the true cells correctly. Across all scenarios, the expected number of correct identifications is no more than 7.5%. As expected, R_{id} is largest for the 30×30 grid. The values of R_{id} for the 10×10 and 20×20 grids are nearly identical, suggesting that most of the identification disclosure risk with these grids comes from cases that have unique combinations of (i, b) across all grids. These R_{id} values are smaller than R_{id} for the 30×30 grid since larger areas contain fewer unique combinations of (i, b).

Figure 4.10 displays histograms of d_t for all n individuals. In the 10×10 grid, a distance greater than 2.12 means that s_t is not one of the neighbors of the intruder's best guess. This threshold is 1.06 and .71 for the 20×20 and 30×30 grids, respectively.

Table 4.2: Summary of risk measures for the three grid sizes.

Grid size	10×10	20×20	30×30
R_{all}	.082	.060	.088
R_{unq}	.134	.080	.112
R_{id}	.045	.044	.074

As evident in Figure 4.10, in each scenario most d_t exceed these critical distances. Thus, the intruder's guesses tend to be far from the true locations.

4.5 Concluding Remarks

The results from the synthesis of the Durham, NC, mortality data illustrate how tuning the grid size effectively trades disclosure risk for data quality. Given that the risk computations presumed an intruder with very, and perhaps unrealistically, detailed knowledge, we suspect that many agencies would be comfortable with the risks of releasing the synthetic data generated via the 20×20 or 30×30 grids. These releases were superior in data quality compared to the synthetic data generated via the 10×10 grid. Of course, agencies can and should evaluate the quality of additional representative statistical analyses when comparing the risk-utility profiles of any proposed release, as well as consider (when sensible) multiple grids of differing sizes.

After generating synthetic locations, the agency still may deem the disclosure risks too large for some records. As suggested by a reviewer, *post hoc* agencies can smooth the synthetic location probabilities for each risky case over additional grid cells, and re-draw synthetic locations for those cases. Alternatively, agencies can use the location probabilities from coarser synthesis models for the risky cases. When the number of risky cases is small, either of these two approaches should not seriously degrade the quality of the synthetic data. Such *post hoc* changes in

the synthesis model should be reflected (at least approximately) in the likelihood $P(\tilde{S}|s_t = s, X, A, M)$ when re-computing the risk measures.

We simulated locations only for Durham, NC, which comprised 6294 cases. Extending to much larger data sets, for example the entire state of NC, demands more efficient computational algorithms than those described and used here. One convenient strategy is to partition the data into geographical regions effectively modeled with manageable grid size, and simulate grid cells by running the synthesizer independently within each region. This has the added benefit of exactly preserving spatial analyses that use the regions as the finest level of geography. One also could tailor the grid size in each region to improve risk-utility profiles, for example using a coarser grid in regions where disclosure risks appear to be high and a finer grid in regions where risks appear to be low.

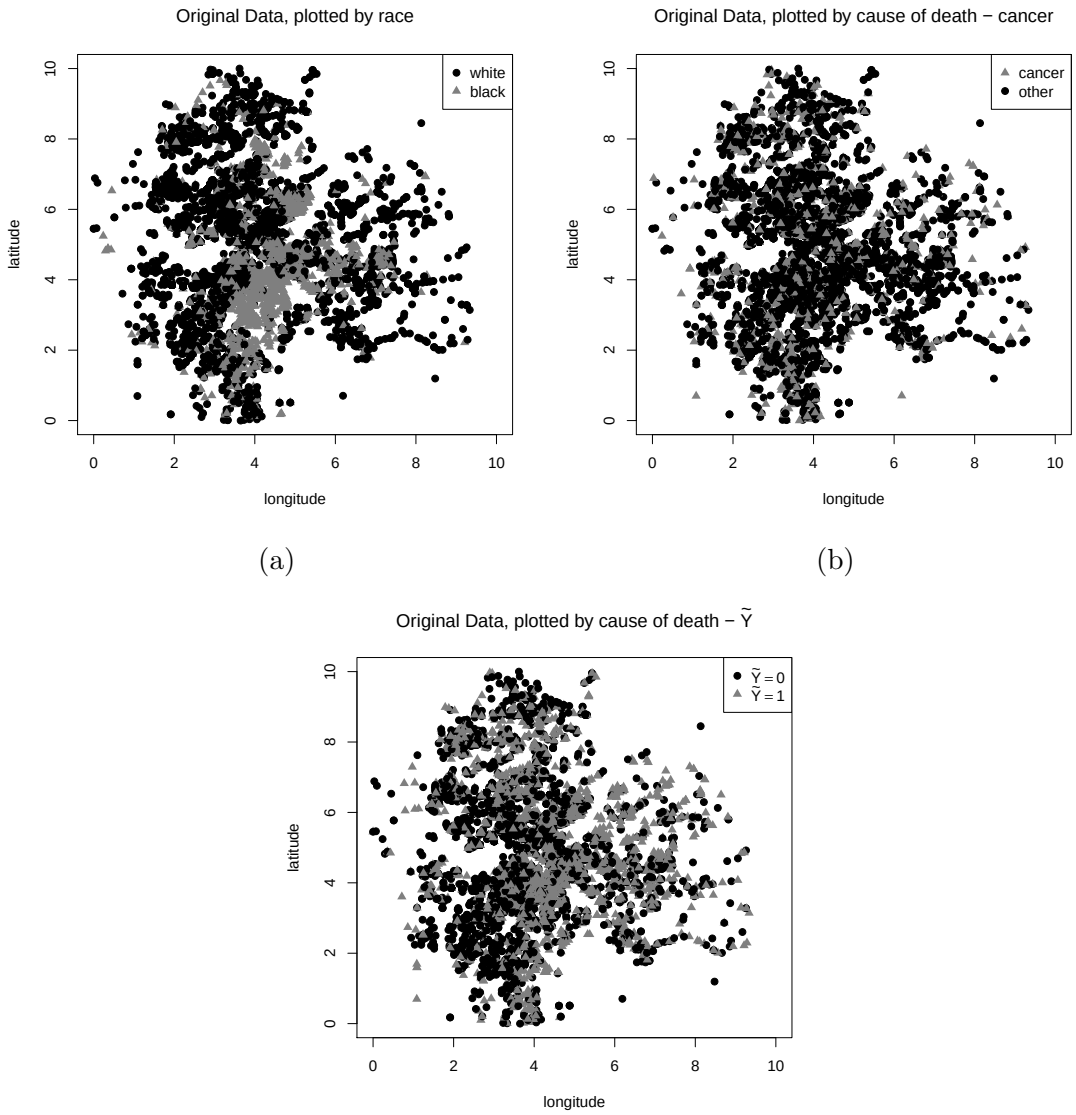
The synthetic data reflect only the relationships between geographies and attributes that are encoded in the synthesis models. Thus, non-geographic attributes not in the models may have distorted synthetic spatial distributions. Similar problems arise when adding non-geographic attributes from another database to the original file by means of matching the actual locations from the other database to the synthetic locations. The synthesized geographies are conditionally independent of the appended attributes (given the attributes in the original file), which may not mimic reality. On the other hand, synthetic geographies can potentially allow analysts to discern associations between the attributes in the original file and contextual variables affiliated with geographies (at the grid cell or coarser levels), such as the number of parks or grocery stores near a particular location, even if the contextual variables derive from external information. For example, suppose a disease occurs most often in a particular set of grid cells marked by an unusual feature, such as heavy pollutors. When disease status is included in the model, the synthetic data should appropriately impute that set of locations for people with the disease, and

hence connect the disease incidence to the locations of the pollutors.

These issues suggest that agencies include in the synthesis model as many attributes that vary spatially (as indicated by exploratory data analyses and prior scientific knowledge) as possible while ensuring acceptable disclosure risks. The overarching goal of model selection is to reflect the spatial relationships in the collected sample faithfully as opposed to, for example, minimizing out-of-sample prediction errors. When the number of potentially relevant variables is large enough to make model estimation unwieldy, the agency can exclude attributes that do not make important contributions to the predicted Poisson rates.

The agency can evaluate the synthetic geography models by comparing inferences made with synthetic locations to those made with observed locations, using analyses representative of those anticipated to be of interest to users (Reiter and Drechsler, 2010). The agency can release such evaluations to the public so that analysts can assess what types of analyses are not supported by the synthetic geographies. Another possibility is for agencies to provide feedback to analysts about the quality of the synthetic data inferences for specific estimands; see Reiter et al. (2009) and McClure and Reiter (2012) for proposals to build automated systems that offer such feedback.

Disease mapping models disregard within-grid cell heterogeneity in spatial intensity surfaces. With fine enough grids, in many data settings such heterogeneity may be modest enough to be swamped by the inherent variability in the synthesis process. When preservation of finer spatial structure is desired, one could use log Gaussian Cox processes (Møller et al., 1998) and associated computational strategies for fitting them (Brix and Diggle, 2001; Rodrigues and Diggle, 2012). We leave comparative evaluation of point pattern models and disease mapping models on dimensions of risk, utility, and computational expediency for future research.



(a) (b)

(c)

FIGURE 4.1: Plots of original locations labeled by different attributes

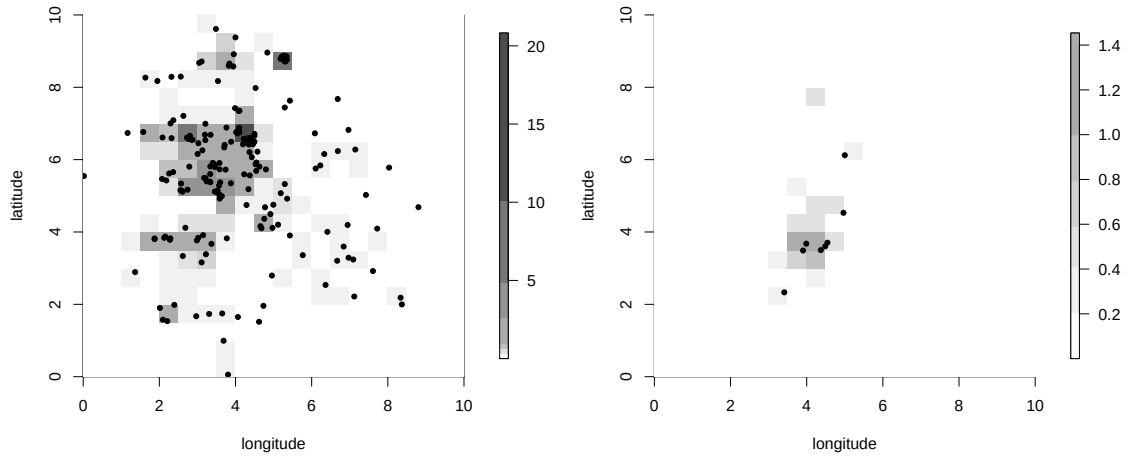


FIGURE 4.2: Posterior mean surface of λ for the 20×20 grid for white women over age 85 with education less than high school and $\tilde{Y} = 0$ (left), and for black men less than age 60 with more than four years of college and $\tilde{Y} = 1$ (right).

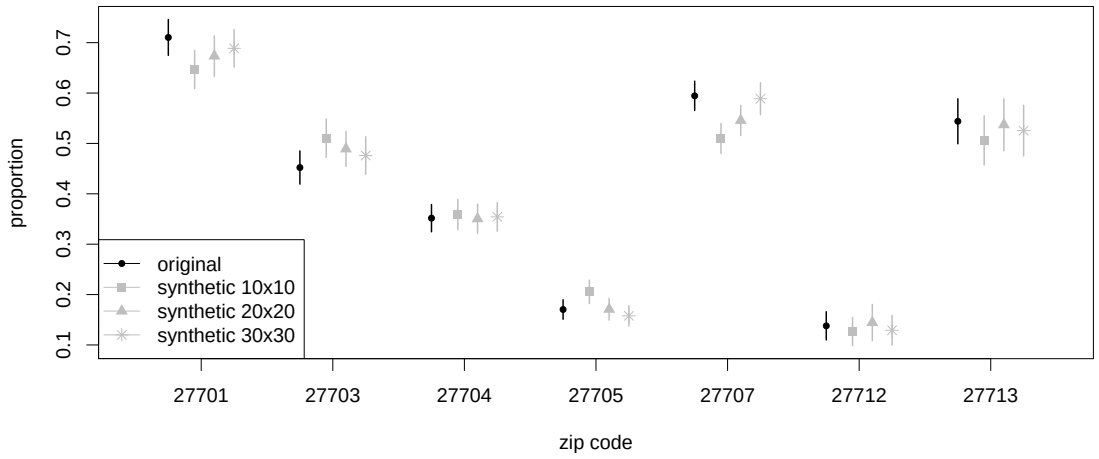


FIGURE 4.3: Comparison of point estimates and 95% confidence intervals for proportion of black people per zip code, estimated with the original and synthetic data for the three grid sizes.

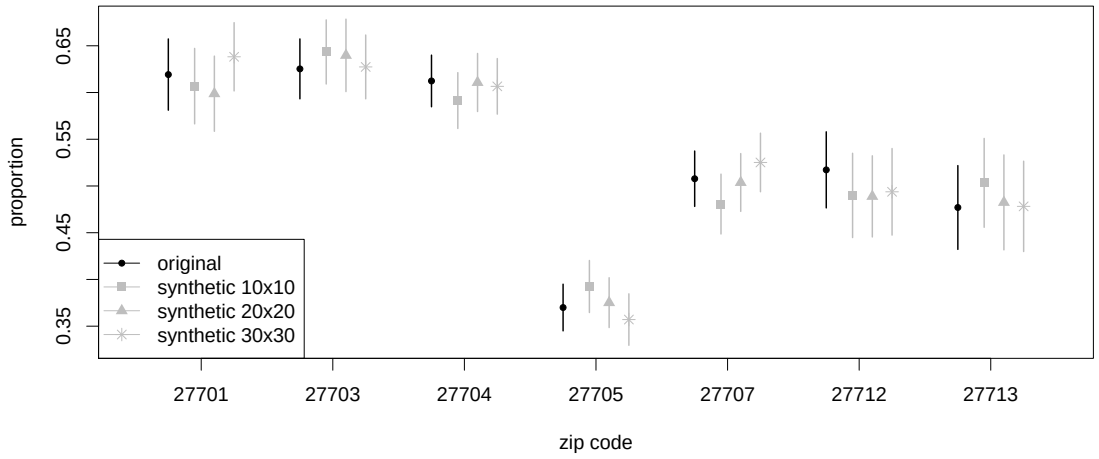


FIGURE 4.4: Comparison of point estimates and 95% confidence intervals for the proportion of $\tilde{Y} = 1$ per zip code, estimated with the original and synthetic data for the three grid sizes.

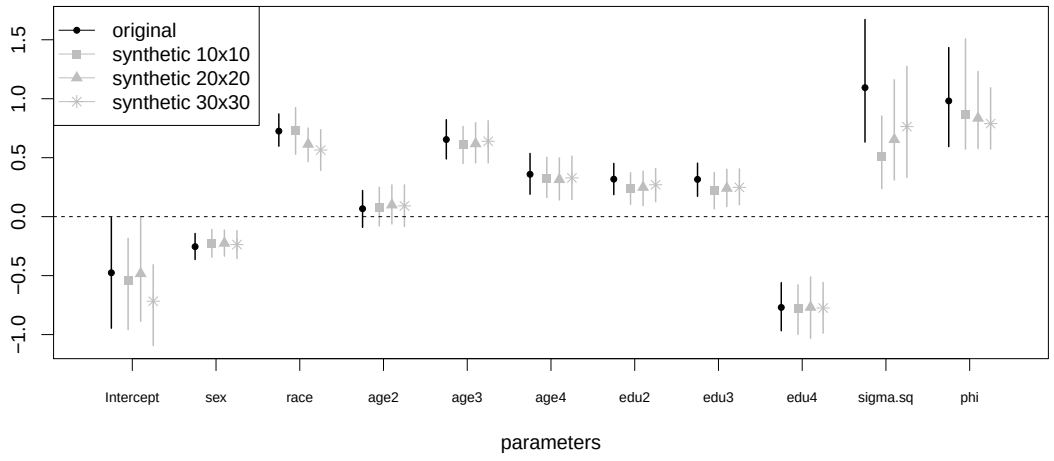


FIGURE 4.5: Comparison of posterior means and 95% credible intervals from the spatial regression, estimated with the original and the synthetic data sets for the three grid sizes.

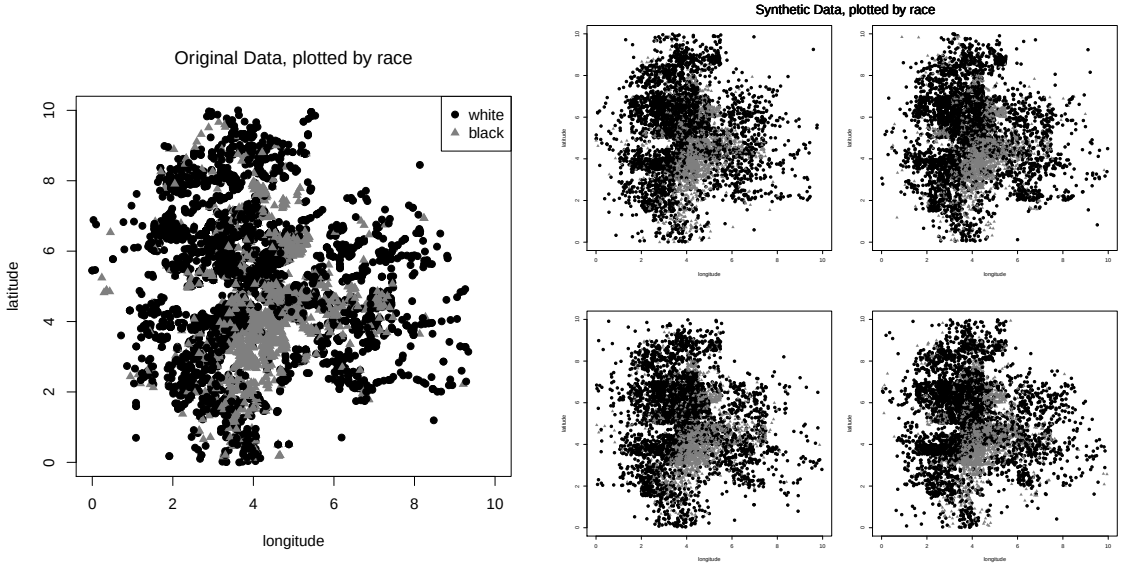


FIGURE 4.6: Plot of the original and synthetic locations labeled by race

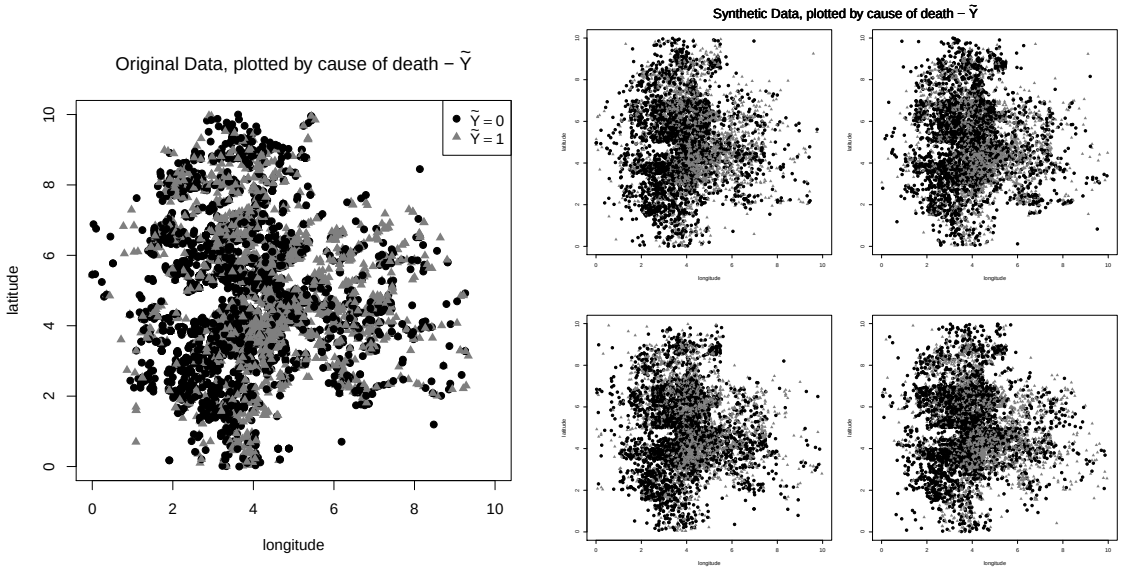


FIGURE 4.7: Plot of the original and synthetic locations labeled by cause of death $\tilde{Y}$

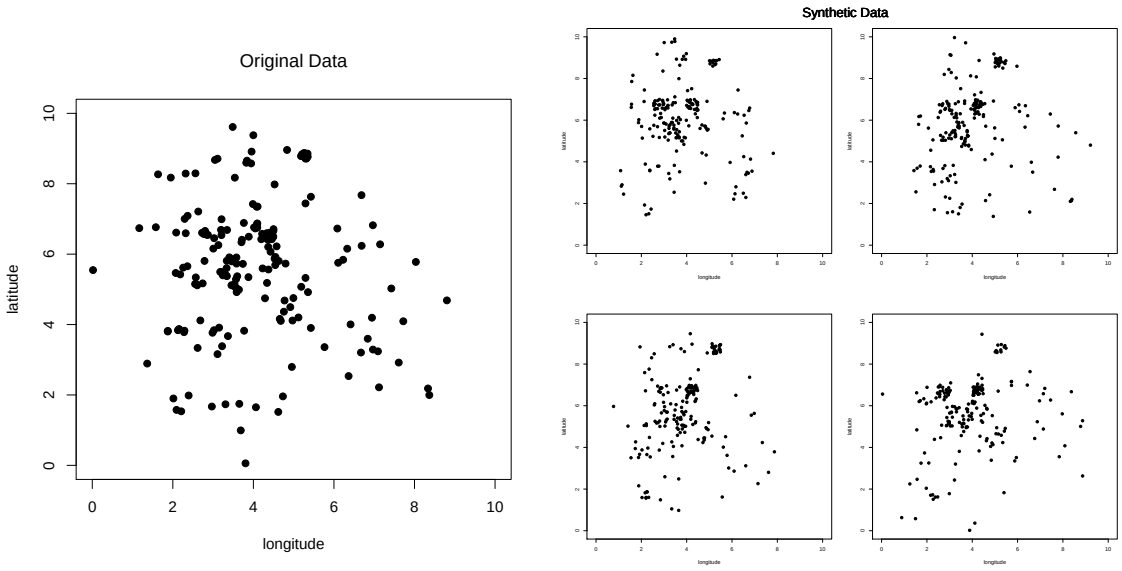


FIGURE 4.8: Plots of the original and synthetic locations for white women over age 85 with education less than high school and $\tilde{Y} = 0$.

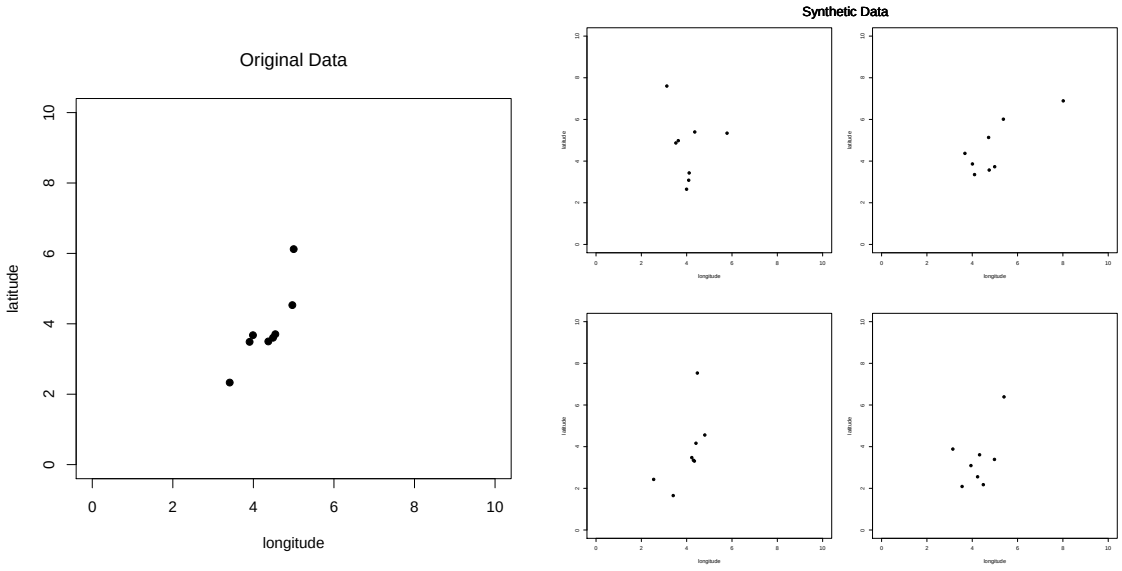


FIGURE 4.9: Plots of the original and synthetic locations for black men less than age 60 with more than four years of college and $\tilde{Y} = 1$.

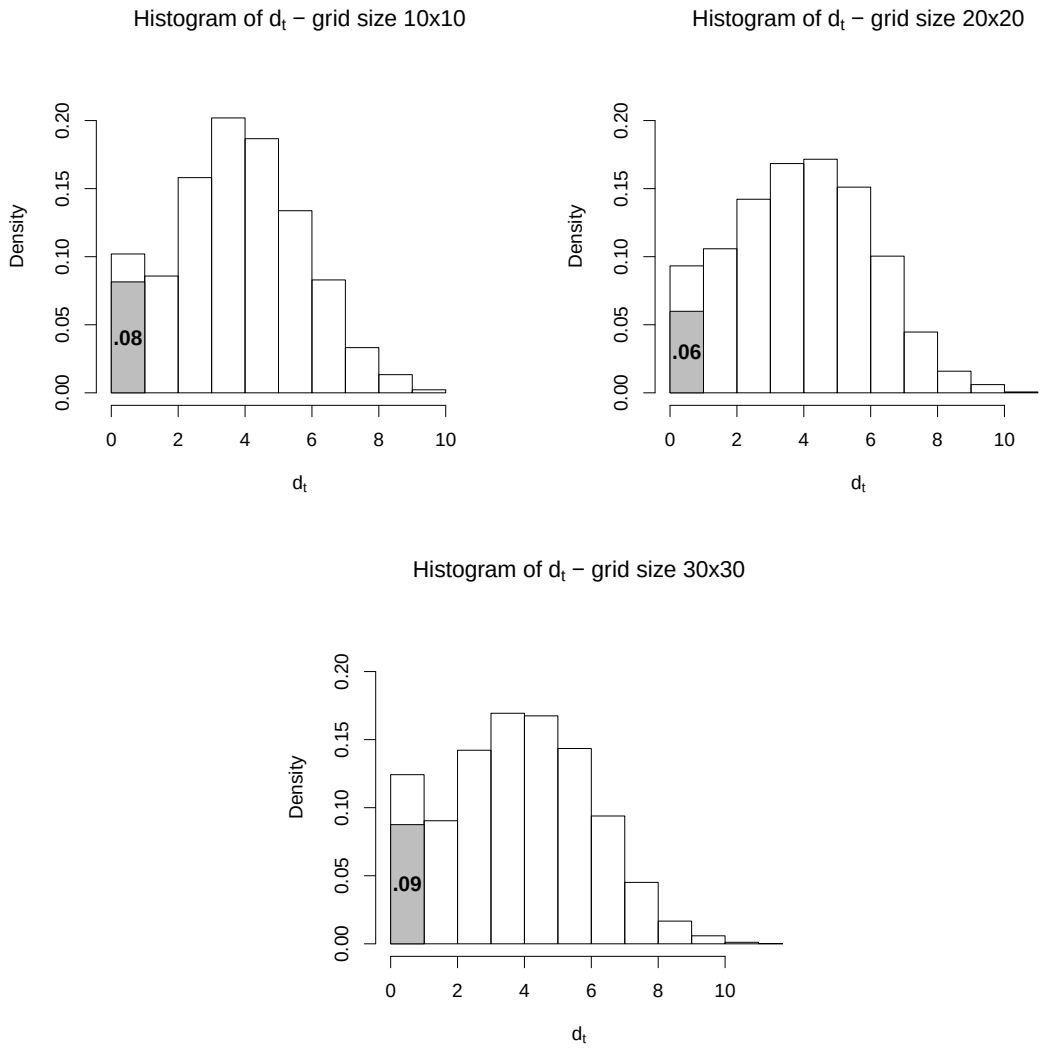


FIGURE 4.10: Histogram of d_t, for the different grid sizes. The highlighted area represents the proportion of observations with $r_t = 1$.

Bibliography

An, D., Little, R., and McNally, J. W. (2010). A multiple imputation approach to disclosure limitation for high-age individuals in longitudinal studies. *Statistics in Medicine*, 29:1769–1778.

Armstrong, M. P., Rushton, G., and Zimmerman, D. L. (1999). Geographically masking health data to preserve confidentiality. *Statistics in Medicine*, 18(5):497–525.

Banerjee, S., Gelfand, A. E., and Carlin, B. P. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC Press, Boca Raton.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848.

Barnard, J. and Meng, X.-L. (1999). Applications of multiple imputation in medical studies: from aids to nhanes. *Statistical Methods in Medical Research*, 8(1):17–36.

Barnard, J. and Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4):pp. 948–955.

Besag, J., York, J., and Molli, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43:1–20.

Brick, J. and Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5(3):215–238.

Brix, A. and Diggle, P. J. (2001). Spatiotemporal prediction for log-Gaussian Cox processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):823–841.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43(3):pp. 671–681.

Clayton, D. G. and Bernardinelli, L. (1992). Bayesian methods for mapping disease risk. In Elliott, P., Cuzick, J., English, D., and Stern, R., editors, *Geographical and Environmental Epidemiology: Methods for Small Area Studies*, pages 205–220. Oxford University Press, Oxford; New York.

Daniels, M. and Hogan, J. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Chapman & Hall/CRC Press. Taylor & Francis, Boca Raton.

Daniels, M. J. and Hogan, J. W. (2000). Reparameterizing the pattern mixture model for sensitivity analyses under informative dropout. *Biometrics*, 56(4):1241–1248.

Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied statistics*, 43(1):49–93.

Drechsler, J. and Reiter, J. P. (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In Domingo-Ferrer, J. and Saygin, Y., editors, *Privacy in Statistical Databases (LNCS 5262)*, pages 227–238. New York: Springer-Verlag.

Duncan, G. T. and Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics*, 7:207–217.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. *Recent advances in statistics*, 24:287–302.

Fienberg, S. E., Makov, U. E., and Sanil, A. P. (1997). A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data. *Journal of Official Statistics*, 13:75–89.

Finamore, J., Reist, B., and Coffey, S. (2013). 2013 National survey of college graduates: A practice-based investigation of adaptive design. In *AAPOR (American Association for Public Opinion Research) 68th Annual Conference*.

Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009). Improving the performance of predictive process modeling for large datasets. *Computational Statistics & Data Analysis*, 53(8):2873 – 2884.

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2008). *Longitudinal Data Analysis*. Chapman & Hall/CRC. Taylor & Francis, Boca Raton.

Forster, J. and Webb, E. (2007). Bayesian disclosure risk assessment: predicting small frequencies in contingency tables. *Journal of the Royal Statistical Society: Series C*, 56:551 – 557.

Fraley, C. and Raftery, A. E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, 24(2):155–181.

Gelfand, A. E., Holder, M., Latimer, A., Lewis, P. O., Rebelo, A. G., Silander, J. A., and Wu, S. (2006). Explaining species distribution patterns through hierarchical modeling. *Bayesian Analysis*, 1(1):41–92.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(2):337–348.

Glynn, R., Laird, N., and Rubin, D. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In Wainer, H., editor, *Drawing Inferences from Self-Selected Samples*, pages 115–142. Springer New York.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1993). Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. *Journal of the American Statistical Association*, 88(423):pp. 984–993.

Graham, J. (2012). *Missing Data: Analysis and Design*. Springer, New York.

Greenlees, J. S., Reece, W. S., and Zieschang, K. D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77(378):pp. 251–261.

Groves, R. (2004). *Survey Errors and Survey Costs*. Wiley Series in Probability and Statistics. Wiley, New York.

Groves, R. M. and Heeringa, S. G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3):439–457.

Harel, O. and Zhou, X.-H. (2007). Multiple imputation: review of theory, implementation and software. *Statistics in medicine*, 26(16):3057–3077.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453).

Kalton, G. and Kasprzyk, D. (1986). Treatment of missing survey data. *Survey Methodology*, 12:1–16.

Kaufman, C. G. and Sain, S. R. (2010). Bayesian functional ANOVA modeling using Gaussian process prior distributions. *Bayesian Analysis*, 5(1):123–149.

Kim, H. J., Reiter, J. P., Wang, Q., Cox, L. H., and Karr, A. F. (2014). Multiple imputation of missing or faulty values under linear constraints. *Journal of Business & Economic Statistics*, 32(3):375–386.

Li, K. H., Raghunathan, T. E., and Rubin, D. B. (1991). Large-sample significance levels from multiply imputed data using moment-based statistics and an f reference distribution. *Journal of the American Statistical Association*, 86(416):pp. 1065–1073.

Little, R. (2008). Selection and pattern-mixture models. In Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G., editors, *Longitudinal Data Analysis*, pages 409–431. Chapman and Hall/CRC, Boca Raton.

Little, R. and Rubin, D. (1987). *Statistical Analysis With Missing Data*. Wiley Series in Probability and Statistics. Wiley, New York.

Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, Second edition.

Little, R. J. (1993a). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):pp. 125–134.

Little, R. J. (1993b). Statistical analysis of masked data. *Journal of Official Statistics*, 9:407–407.

Little, R. J. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81(3):471–483.

Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431):1112–1121.

Liu, J., Gelman, A., Hill, J., Su, Y.-S., and Kropko, J. (2014). On the stationary distribution of iterative imputations. *Biometrika*, 101(1):155–173.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008). Privacy: Theory meets practice on the map. In *IEEE 24th International Conference on Data Engineering*.

Manrique-Vallier, D. and Reiter, J. P. (2012). Estimating identification disclosure risk using mixed membership models. *Journal of the American Statistical Association*, 107(500):1385–1394.

McClure, D. R. and Reiter, J. P. (2012). Towards providing automated feedback on the quality of inferences from synthetic datasets. *Journal of Privacy and Confidentiality*, 4(1):8.

Meng, X.-L. and Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, 79(1):pp. 103–111.

Miller, P. V. (2013). Adaptive design at the Census Bureau - a new way of doing business. In *AAPOR (American Association for Public Opinion Research) 68th Annual Conference*.

Molenberghs, G. (2009). Incomplete data in clinical studies: analysis, sensitivity, and sensitivity analysis. *Drug Information Journal*, 43(4):409–429.

Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):371–388.

Molenberghs, G. and Kenward, M. (2007). *Missing Data in Clinical Studies*. Statistics in Practice. Wiley, Chichester; Hoboken, NJ.

Molenberghs, G., Kenward, M. G., and Lesaffre, E. (1997). The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika*, 84(1):33–44.

Molenberghs, G., Michiels, B., Kenward, M. G., and Diggle, P. J. (1998). Monotone missing data and pattern-mixture models. *Statistica Neerlandica*, 52(2):153–161.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482.

Müller, P. and Mitra, R. (2013). Bayesian nonparametric inference – why and how. *Bayesian Analysis*, 8(2):269–302.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355.

National Research Council (2005). *Expanding Access to Research Data: Reconciling Risks and Opportunities*. Panel on Data Access for Research Purposes, Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington DC: The National Academies Press.

National Research Council (2007). *Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data*. Washington DC: The National Academies Press.

Paiva, T., Chakraborty, A., Reiter, J., and Gelfand, A. (2014). Imputation of confidential data sets with spatial locations using disease mapping models. *Statistics in Medicine*, 33(11):1928–1945.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27(1):85–96.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19(1):1–16.

Rao, R. S., Glickman, M. E., and Glynn, R. J. (2008). Stopping rules for surveys with multiple waves of nonrespondent follow-up. *Statistics in medicine*, 27(12):2196–2213.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29(2):181–188.

Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, 30(235):l242.

Reiter, J. P. (2005). Estimating identification risks in microdata. *Journal of the American Statistical Association*, 100:1103–1113.

Reiter, J. P. (2007). Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika*, 94(2):502–508.

Reiter, J. P. and Drechsler, J. (2010). Releasing multiply-imputed, synthetic data generated in two stages to protect confidentiality. *Statistica Sinica*, 20:405–422.

Reiter, J. P. and Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality*, 1:99–110.

Reiter, J. P., Oganian, A., and Karr, A. F. (2009). Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Computational Statistics and Data Analysis*, 53:1475–1482.

Reiter, J. P. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102(480):1462–1471.

Robert, C. P. and Casella, G. (2005). *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Rodrigues, A. and Diggle, P. J. (2012). Bayesian estimation and prediction for inhomogeneous spatiotemporal log-Gaussian Cox processes using low-rank models, with application to criminal surveillance. *Journal of the American Statistical Association*, 107(497):93–101.

Rodríguez, C. E. and Walker, S. G. (2014). Label switching in Bayesian mixture models: Deterministic relabeling strategies. *Journal of Computational and Graphical Statistics*, 23(1):25–45.

RStudio and Inc. (2014). *shiny: Web Application Framework for R*. R package version 0.9.1.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, Hoboken, NJ.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, 72(359):pp. 538–543.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9(2):461–468.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489.

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, London; New York.

Schafer, J. L. and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, 7(2):147.

Schafer, J. L. and Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate behavioral research*, 33(4):545–571.

Scharfstein, D. O., Daniels, M. J., and Robins, J. M. (2003). Incorporating prior beliefs about selection bias into the analysis of randomized trials with missing outcomes. *Biostatistics*, 4(4):495–512.

Schouten, B., Bethlehem, J., Beullens, K., Kleven, ., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N., and Skinner, C. (2012). Evaluating, comparing, monitoring, and improving representativeness of survey response through r-indicators and partial r-indicators. *International Statistical Review*, 80(3):382–399.

Schouten, B., Calinescu, M., and Luiten, A. (2013). Optimizing quality of response through adaptive survey designs. *Survey Methodology*, 39(1):29–58.

Schouten, B., Cobben, F., and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35(1):101–113.

Schouten, B., Shlomo, N., and Skinner, C. (2011). Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics*, 27(2):1–24.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.

Shlomo, N., Skinner, C., and Schouten, B. (2012). Estimation of an indicator of the representativeness of survey response. *Journal of Statistical Planning and Inference*, 142(1):201 – 211.

Skinner, C. (2012). Statistical disclosure risk: Separating potential and harm. *International Statistical Review*, 80:349 – 368.

Skinner, C. J. and Shlomo, N. (2008). Assessing identification risk in survey microdata using log-linear models. *Journal of the American Statistical Association*, 103:989–1001.

Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., and Curran, D. (2002). Strategies to fit pattern-mixture models. *Biostatistics*, 3(2):245–265.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3). Open Access.

Van Buuren, S. and Oudshoorn, C. (2000). *Multivariate Imputation by Chained Equations: MICE V1. 0 Users's Manual*. TNO Prevention and Health, Public Health.

VanWey, L. K., Rindfuss, R. R., Guttman, M. P., Entwisle, B., and Balk, D. L. (2005). Confidentiality and spatially explicit data: concerns and challenges. *Proceedings of the National Academy of Sciences*, 102:15337–15342.

Wagner, J. (2008). *Adaptive Survey Design to Reduce Nonresponse Bias*. PhD thesis, University of Michigan.

Wagner, J. and Raghunathan, T. E. (2010). A new stopping rule for surveys. *Statistics in medicine*, 29(9):1014–1024.

Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, 8(2):158–183.

Wang, H. and Reiter, J. P. (2012). Multiple imputation for sharing precise geographies in public use data. *Annals of Applied Statistics*, 6(1):229–252.

West, M., Müller, P., and Escobar, M. D. (1994). Hierarchical priors and mixture models, with applications in regression and density estimation. *Aspects of Uncertainty: A Tribute to D. V. Lindley*, pages 363–386.

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.

Woo, M.-J., Reiter, J. P., and Karr, A. F. (2008). Estimation of propensity scores using generalized additive models. *Statistics in Medicine*, 27(19):3805–3816.

Woo, M.-J., Reiter, J. P., Oganian, A., and Karr, A. F. (2009). Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*, 1(1):7.

Young, C., Martin, D., and Skinner, C. J. (2009). Geographically intelligent disclosure control for flexible aggregation of census data. *International Journal of Geographical Information Science*, 23:457–482.

Zayatz, L. (2007). Disclosure avoidance practices and research at the U. S. Census Bureau: an update. *Journal of Official Statistics*, 23:253–255.

Zhou, X. and Reiter, J. P. (2010). A note on Bayesian inference after multiple imputation. *The American Statistician*, 64(2):159–163.

Zhou, Y., Dominici, F., and Louis, T. A. (2010). A smoothing approach for masking spatial data. *Annals of Applied Statistics*, 4(3):1451–1475.

Biography

Thais Viana Paiva was born in December 1, 1987, in Belo Horizonte, Minas Gerais, Brazil. She attended the Federal University of Minas Gerais (UFMG) in Belo Horizonte, where she received her B.S. degree in Actuarial Science in December, 2008, and her M.S. in Statistics in August, 2010. After that, she attended Duke University in Durham, NC, USA, where she received her Masters *en route* to her Ph.D. in May, 2012, and plans to graduate with her Ph.D. in October, 2014, under the supervision of Professor Jerome Reiter. After graduating, she will be working as a postdoctoral scholar with Professor Renato Assunção, back in the Federal University of Minas Gerais, in Belo Horizonte, Brazil.